Effective altruism, risk, and human extinction

Richard Pettigrew (University of Bristol)

Global Priorities Institute | July 2022

GPI Working Paper No . 2-2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Longtermism and risk

Richard Pettigrew

July 13, 2022

Abstract

A future in which humanity does not go extinct from something like a meteor strike or nuclear war might contain vast quantities of great happiness and human flourishing. But it might also contain vast quantities of great misery and wasted potential. Any effort to ensure a long happy future will have to strive both to ensure a long future, by reducing extinction risks, and to ensure it is a happy one, by improving the quality of life. Such an effort might succeed at the former goal without succeeding at the latter. So any effort to ensure a long happy future will increase the chance of such a future, but it will also increase the chance of a long miserable future, even if it increases the latter by a smaller amount. Granted this, if you are risk-averse, or if morality requires you to choose using risk-averse preferences, then you do better to work to bring about humanity's extinction than to secure its long-term survival. This conclusion will seem troubling to many, though perhaps welcome to some. In this paper, I try to formulate the argument as plausibly and as robustly as I can. I then investigate how those who wish to resist it might do so.

1 Introduction

Let me begin on the final few pages of *Reasons and Persons*, where Parfit asks us to consider three possibilities (Parfit, 1984, 453-4):

(1) Peace;

(2) A nuclear war that kills 99% of the human population;

(3) A nuclear war that kills 100% of the human population.

He claims that (1) is better than (2), and (2) is better than (3). Most people, he thinks, would agree to this. Then he asks by how much (1) is better than (2), and by how much (2) is better than (3). He thinks the difference between (2) and (3) is vastly greater than the difference between (1) and (2). His reason? Barring a major extinction event such as a nuclear winter, a dramatically virulent disease, or a massive meteor strike, Earth will remain

inhabitable for humans for a further billion years.[1] We might suppose that, in scenario (1), we will go on to inhabit the planet for that whole period, and at a stable population of 10 billion humans at any given time. In that case, scenario (1) will contain 10 billion billion years of human life. In scenario (2), while the population will drop from 8 billion to 80 million initially, we might suppose that, after at most a thousand years, it will bounce back and humanity will continue for the remainder of the billion years at the stable population of 10 billion. So the difference between (1) and (2) is a little under 8,000 billion years of human life—the life years lost during the millenium of recovery following the war. This is an enormous difference by any reckoning. But now consider the difference between (2) and (3). Scenario (3) contains no further years of human life. And so the difference between (2) and (3) is the 10 billion billion years of human life that (1) contains less the 8,000 billion life years lost to the war, which is more than 9 billion billion years of human life. So, as Parfit says, the difference between (2) and (3)—more than 9 billion billion years of human life—is vastly greater than between (1) and (2)—just under 8,000 billion life years. Indeed, more than a million times greater.

Parfit himself doesn't do much to draw out the consequences of this insight. But, in recent years, some philosophers within the Effective Altruism movement, known as *longtermists*, have noted that it has obvious implications for certain sorts of consequentialist (Bostrom, 2013; Beckstead, 2013; Ord, 2020; Greaves & MacAskill, 2021). Simplifying greatly, suppose I must choose to which of two charities I will donate a certain sum of money. If I donate it to the Against Malaria Foundation (AMF), which distributes insecticidal nets in malarial regions of the world, there is a 90% chance that it will save the lives of ten children under five years old who would otherwise have died from malaria and who will now go on to live for a further 70 years; and there is a 10% chance that it will have no effect. So, were I to donate to the AMF, it would, in expectation, increase the number of human life years by 630.[2] On the other hand, if I donate to the Longtermism Fund (LF), which works to reduce the chance of a catastrophic event that results in human extinction, I will lower by one in a millon billion the chance of Parfit's third scenario (3)—and thereby raise by one in a million billion the chance of his first two scenarios, (1) or (2). So a donation to the LF will result in around 9,000 expected life years.[3] Then, the argument goes, we

[1] Here, I use the estimate from Hilary Greaves and William MacAskill's work, though the point is easily made with very different estimates (Greaves & MacAskill, 2021). Since one upshot of this paper is that there are problems with their arguments and the arguments of other longtermists, I try to use reasonably similar assumptions where I can.

[2] 10 children, each with 70 extra years of life, gives 700 life years. And then 90% of that is 630 life years.

[3] At least 9 billion billion (9×10^{18}) life years in scenario (1) or (2), and an increased chance of one in 1 million billion (10^{-15}), which gives at least 9,000 expected life years.

should donate in the way that will maximise the expected number of life years, and that means we should donate to LF when the alternative is AMF.

In this paper, I wish to raise a worry about this argument. I'll begin by improving it to ensure that we are considering the best version. Like the version just given, the improved argument relies on a particular consequentialist account of how to determine which is the morally correct or morally best option to choose when faced with a decision between a range of available options, such as the decision between donating to AMF or to LF or not donating at all. This account relies on applying rational choice theory in ethics. Rational choice theory is typically used to determine which are the rationally permissible options to choose when faced with a decision. Typically, in order to make that determination, it asks for some ingredients: it asks you to enumerate a set of possible states of the world, each of which specifies, up to a certain level of grain, how things have gone in the past, how they are now, and how they will go in the future; it asks you to assign to each option you might choose and each of these states of the world a subjective utility that measures how much you personally would value the outcome of choosing that option should the world be in that state; and it asks you to assign to each possible option and each state of the world a probability that measures how likely you personally think it is that the state of the world will come about should you choose that option. A theory of rational choice then offers a recipe by which to combine these various components to give an ordering of the options in terms of their subjective choiceworthiness, and many of them then say that rationality requires you to choose from among the options at the top of the ordering.[4] So, for instance, expected utility theory orders options by their expected utility, where the expected utility of an option is the sum of the utilities of its possible consequences, each weighted by the probability that the consequence will come about given you choose that option. And it requires you to choose an option from the top of the ordering.

When consequentialist arguments of the sort considered here apply rational choice theory in ethics, they switch out some of these ingredients. They typically retain the states of the world you specified, as well as your subjective probabilities, providing the latter satisfy all constraints of epistemic rationality, such as obeying the axioms of the probability calculus, being proportioned to your evidence, and so on. But they replace your subjective utilities—which, recall, measure how much you personally value a state of the world—with something like morally ideal utilities—which measure either how valuable a state of the world is from a completely impersonal point of view, or how much you would value that state of the

[4]Others, which don't assume that the ordering is total, might say you should choose among those options for which there is no greater option in the ordering; and others still, which are less demanding, say you should choose among those options that lie above a certain point on the ordering.

world were you to give at least the morally required level of consideration to the lives of all others, which may not be the same level as your nearest and dearest, but which is nonetheless higher than most people give. Having tweaked the ingredients in this way, the consequentialist then appeals to the same recipe by which to combine them to give an ordering of the options in terms of their moral choiceworthiness, and they say that morality requires you to choose an option at the top. In other words, moral decision making is just rational decision making with the utilities that measure the impersonal good, or at least a more impersonal good than most people actually value.

For instance, take the consequentialist argument for donating to LF instead of AMF. The impersonal or morally required conception of the good is simply calculated by counting up the number of human life years lived; and the theory of rational choice to which the argument appeals is expected utility theory. The improved version of the argument I will give below appeals to a more plausible account of the impersonal good, but it retains the appeal to expected utility theory. The problem with this argument that I wish to explore is that expected utility theory is not the correct theory of rational choice. According to the correct theory, the ingredients of rational choice are not only states of the world, probabilities, and utilities; rather, they are joined by a specification of the decision-maker's attitudes to risk. And then the probabilities, utilities, and attitudes to risk are combined in a particular way to give the ordering of choiceworthiness from which we are rationally required to choose an option at the top. Now, for certain decision-makers—namely, those who are risk-neutral in a particular sense—this recipe returns the same ordering that expected utility theory does. But for other decision-makers—namely, those who are risk-averse or risk-inclined—it does not. So, at the most, the argument for donating to LF applies only to risk-neutral individuals, and experimental evidence suggests that such individuals are rather rare. But, in fact, I'll go further. I'll argue that, when we assess the moral correctness of a decision, we should not use the decision maker's subjective attitudes to risk, but instead an aggregate of the attitudes to risk endorsed by the individuals affected by the decision-maker's choice; and, moreover, this aggregate should place more weight on the risk-averse individuals. Thus, when we assess the moral correctness of a decision, we will typically be required to use a theory of rational choice that governs risk-averse decision-makers. When we replace expected utility theory with such a theory of rational choice in the argument for donating to LF, that argument no longer goes through. Indeed, not only does the argument for working to avoid our extinction no longer work; it is instead transformed into an argument for working to hasten that extinction.[5]

[5]In her 2019 Parfit Memorial Lecture at the Global Priorities Institute, Lara Buchak, to

2 The improved argument for working to avoid extinction

The first thing to change in the argument based on Parfit's observation is the measure of impersonal goodness. In that argument, the moral utility function measured the goodness of a state of the world by the number of human life years it contains. But what makes a history good or bad is not only how much or how little human life it contains, but also what that life is like for the people who live it. Lives can go better or worse, be filled more with pleasure than with pain or vice versa; be fulfilling or be thwarted; can allow the subject to perfect their capacities to different degrees or not at all; and so on. For the sake of simplicity, in our new argument, I will assume a straightforward total hedonist utilitarian account of the impersonal good—the goodness of a state of the world is just its total hedonic value, which weighs the amount, intensity, and nature of the pleasure it contains against the amount, intensity, and nature of the pain it contains. That's not because I think this is the correct account; and it's not because I think all proponents of longtermist arguments subscribe to it (though many do). It's simply for concreteness at this stage. My plan is to present a simple, concrete version of the argument first, so that we can see most clearly what happens when we replace the assumption of expected utility theory with a risk-sensitive decision theory. Then, in Section 5.2, when I consider how the longtermist might respond, I'll ask what happens to this argument if we change the conception of the good it uses to include non-human lives and other goods beside lives, to include non-hedonic features of those lives, to include global features of the society and its accomplishments as a whole, rather than merely aggregating those of individuals, and so on.

Of course, once we expand the conception of the impersonal good in this way, we see that there are not merely two possible futures for the world—one in which humanity goes extinct in the near future and one in which it survives for a billion years—but many—one in which humanity soon goes extinct, one in which it survives for a billion years and lives are extremely pleasurable on average, one in which it survives and lives are moderately pleasurable on average, one in which it survives and lives are filled with pain and misery on average, and so on.

Indeed, there are also lots of possible durations for humanity's exis-

whose theory of rational choice I will appeal in this paper, also appeals to that theory to question some related but different conclusions of effective altruist arguments about the long-term future (https://youtu.be/eQWY96f41mU). She uses her theory to argue that, faced with the decision whether to use a given finite pot of money to decrease the chance of extinction or to improve the health of people in the present, we are morally required to try to avert extinction; but faced with the decision whether to use a given finite pot of money to improve the health of people in the present or to improve the chances of a wonderful long-term future, we are morally required to improve the health of people in the present.

tence beyond extinction within a century and survival for a billion years. So the future may contain 10 billion billion extremely pleasurable human life years, 10 billion billion moderately pleasurable ones, 10 million billion moderately pleasurable ones, a thousand billion extremely painful ones, a hundred million lives just worth living, and so on.

Again in the interests of simplicity and concreteness, we will consider just four possible futures. Here they are:

(*A*) *The long happy future*. This is the best-case scenario. Humanity survives for the full billion possible years with a stable population around 10 billion 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.

(*B*) *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 impersonal 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, lives are as happy as in (*A*) on the whole, but humanity fails to avoid all extinction risks, and they go extinct after a thousand years.

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

(*D*) *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.

And here are the three possible options between which you must choose. First, the status quo (SQ), in which you do nothing. Second, you donate to the Quiet End Foundation (QEF), which works to bring about a peaceful, painless end to humanity. Third, you donate to the Happy Future Fund (HFF), which works to ensure a long happy future for the species.

For the moment, I won't try to assign specific utilities to these states of the world nor specific probabilities to their occurrence conditional on choosing an particular option. Instead, I'll assign the following placeholders. That will then let us give a general account of when one option is better than another. When we come to the revised version of the argument later on, we'll have to start feeding in specific numbers—but not yet.

First, the utilities of each possible state of the world:

	A	B	C	D
$U(-)$	w	v	u	$-t$

where $-t < 0 < u < v < w$.[6]

Second, the probabilities of each state of the world given each of the three options, SQ, QEF, and HFF:

	A	B	C	D
$P(-\|\text{SQ})$	p	q	r	s
$P(-\|\text{QEF})$	$p-2\varepsilon$	$q-2\varepsilon$	$r+6\varepsilon$	$s-2\varepsilon$
$P(-\|\text{HFF})$	$p+3\varepsilon$	$q+2\varepsilon$	$r-6\varepsilon$	$s+\varepsilon$

The argument for these assignments runs as follows—inevitably, it is rather speculative. First, suppose you donate to the Quiet End Foundation (QEF), which works to bring about a peaceful, painless end to humanity. Your money isn't going to do much, but it will shift the probabilities a little. Let's say you'll raise the probability of imminent extinction (C) by some amount, say 6ε, where ε is some very very small positive probability, and thereby lower the probability of a future for humanity by the same amount; suppose you'll lower the probability of each of the states (A, B, and D) in which humanity has a future by the same amount, that is, 2ε.

Second, suppose you donate to the Happy Future Fund (HFF), which works to ensure a long happy future for the species. Again, your money isn't going to do much, but again it will shift the probabilities a little. You'll raise the probability of the long happy future (A) by some small amount, say 3ε, and you'll raise the probability of (B) by 2ε. But, and here we come to a crucial point, what HFF does to increase the probability of scenario (A) thereby increases the probability of (D) as well, even if not by as much, since there is always the possibility that their attempts to ensure a long happy future only manage to ensure a long future, and that then opens the possibility of a long miserable future. And indeed many of the ways we might try to ensure a long happy future might make a long miserable one more likely than it would otherwise be. We see something like this with so-called gain-of-function research on viruses (Sharples et al., 2015). In that sort of work, virologists engineer versions of a virus, such as a coronavirus or an influenza virus, that are more dangerous than those that have occurred naturally so far, either by increasing transmissibility or increasing infection fatality rates. They do this in order to understand ways in which the viruses might mutate naturally in the future so that they can develop therapeutics to treat the illnesses caused by these mutations and vaccines to prevent their spread. But doing so of course opens the possibility that these enhanced microbes escape the laboratory where they're created, and

[6] Both expected utility theory and the risk-weighted expected utility theory we'll come to use in the revised version of the argument are insensitive to positive linear transformations of utilities. That is, they order options in exactly the same way whether we measure their value at a state of the world using the utility function $U(-)$ or using a positive linear transformation of it, $\alpha U(-) + \beta$, where $\alpha > 0$. So it is not in fact speaking necessary to specify that the moral utility of D is negative. But it does no harm.

therefore the possibility that a bad actor will exploit them to devastating effect—after all, with only the threat of ever deploying an extremely virulent and deadly virus, someone might hold entire societies to ransom. And the same sort of risk emerges from research into AI safety and nuclear safety. By imagining the most devastating uses to which someone might put these technologies in order to guard against them, we might create blueprints for bad actors and thereby make these uses more likely. So, donating to HFF raises the probability of the long miserable future, let's say by ε. And it decreases the probability of imminent extinction (C) by 6ε.

With these in hand, we can now calculate the expected utility of each option, and see that we should prefer the Happy Future Fund to the status quo, and the status quo to the Quiet End Foundation provided $3u + t < w + v$ and $6u + t < 3w + 2v$. And that's a reasonable assumption: after all, if the difference between the long miserable future and extinction is no more than the difference between the long happy future and extinction, then $t < w$ and $t < 3w$;[7] and if the catch-all good-but-not-great outcome is quite a lot better than the extinction outcome, then $3u < v$ and $6u < 2v$. To avoid cluttering the text with the calculation, I spell it out in Appendix A. So, according to the version of consequentialism that pairs this total hedonist utilitarian axiology with expected utility theory, you should donate to the Happy Future Fund.

3 Rational choice theory and risk

The argument for donating to the Happy Future Fund assumes that expected utility theory is the correct theory of rational choice. But unfortunately it isn't. To get a sense of why it goes wrong, consider the following example.[8] 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

[7] It wouldn't be so hard to challenge this assumption, since some think that the range of pain and misery we can experience is greater than the range of pleasure and fulfilment. If that's right, even expected utility theory might recommend donating to the Quiet End Foundation.

[8] For further motivations, 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).

she must choose between two nature reserves: in one, N1, she's sure to see 49 new species; in the other, N2, 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, where M says that the migration has started and $\overline{M}$ says it hasn't:

	M	$\overline{M}$
N1	49	49
N2	0	100

According to expected utility theory, Sheila should choose to go to N2, since, if each new species adds a single utile to an outcome, N2 has expected utility of 50 utiles, while N1 has 49. And yet it seems quite rational for her to choose N1. 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 N2. If Sheila chooses to go to N1, we might say that she is risk-averse, though perhaps only slightly. N2 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, N1 is a risk-free option: it gives you 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 N2, 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 N2. 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 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 the individual's attitudes 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, $u_1 < u_2 < u_3$. 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

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

Roughly speaking, if R is convex, 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, 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.

To see an example at work, consider Sheila's decision whether to go to reserve N1 or N2. First, the risk-weighted expected utility of the safe option, N1.

$$\begin{aligned}\text{REU}(N1) = (R(P(M \vee \overline{M}|N1) - R(P(\overline{M}|N1)))U(M \,\&\, N1) + \\ R(P(\overline{M}|N1))U(\overline{M} \,\&\, N1) = \\ (1 - R(1/2))49 + R(1/2)49 = 49\end{aligned}$$

since $P(M \vee \overline{M}) = 1$, so $R(P(M \vee \overline{M}|N1)) = 1$. Second, the risk-weighted expected utility of the risky option, N2.

$$\begin{aligned}\text{REU}(N2) = (R(P(M \vee \overline{M}|N2) - R(P(\overline{M}|N2)))U(M \,\&\, N2) + \\ R(P(\overline{M}|N2))U(\overline{M} \,\&\, N2) = \\ (1 - R(1/2))0 + R(1/2)100 = R(1/2)100\end{aligned}$$

So $\text{REU}(N2) < \text{REU}(N1)$ iff $R(0.5) < 0.49$. That is, if Sheila is only a little risk-averse, rational choice theory will require her to choose the safe option, N1.

Now let us apply this to the choice between doing nothing, donating to the Happy Future Fund, and donating to the Quiet End Foundation. In contrast with the expected utility calculation, keeping things general doesn't help a great deal—the devil, in this case, is in the details. So we need to specify some numbers.

First, the utilities that morality requires you to assign—according to the arguments we're discussing, these should measure the impersonal good that a state of the world contains; and, more specifically, we're assuming a total hedonist utilitarian axiology. So first, we must specify a unit. Let's say that each human life year lived with the sort of extraordinary pleasure envisaged in scenario A adds one unit of utility, or utile, to the goodness of the states of the future. So the moral utility of A is 10^{19} utiles, since it

contains 10^{19} human life years at this very high level of utility. And let's say that the moral utility of B is 10^{13} utiles, since it contains 10^{13} life years at that level, or many more life years but at a far poorer level on average, which amounts to the same utility. The moral utility of C is 10^6 utiles, since it contains 100 years lived at the same average level that, in scenario B, when lived for a billion years gave 10^{13} utiles. And finally scenario D. 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 as bad as the greatest pleasure is good. That is, the moral utility of the worst case scenario is simply the negative of the moral utility of the best case scenario, where we are taking our zero point to be the utility of non-existence. So the moral utility of D is -10^{19}.

Second, let's try to assign probabilities to these outcomes in the absence of any intervention on your part; then we'll say how donating to different charities affects these probabilities. Inevitably these assignments will be very speculative. It seems clear that the long mediocre or short happy future 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 in the next century. And I'll say that the long happy future, while very unlikely, is nonetheless much much more likely than the long miserable one. More precisely, I'll say the probability of extinction C is one in a thousand ($\frac{1}{10^3}$), the long future A is a thousand times less likely than that ($\frac{1}{10^6}$), the long miserable future D is a thousand times less likely than that ($\frac{1}{10^9}$), and the long mediocre or short happy future mops up the rest of the probability ($1 - \frac{1}{10^6} - \frac{1}{10^3} - \frac{1}{10^9}$).

Finally, as before, let's consider three interventions you might make: do nothing (SQ), donate to the Quiet End Foundation (QEF), or donate to the Happy Future Fund (HFF). As before, your money isn't going to do much, but it will shift the probabilities a little. Let's assume that the interventions change the probabilities in the way we described above, but with $\varepsilon = \frac{1}{10^{10}}$. So, if you donate to QEF, you'll raise the probability of imminent extinction (C) by $\frac{6}{10^{10}}$, and so on.

Once again, I leave the calculations to Appendix A and here only report the conclusion. Suppose your risk function is $R(x) = x^n$ for $n > \frac{3}{2}$. 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. Then $\text{REU}(\text{HFF}) < \text{REU}(\text{SQ}) < \text{REU}(\text{QEF})$. 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 will be to donate to charities that work towards a peaceful end to humanity.

For many, this in itself is a reductio of the approach we've been exploring. For them, it can never be the morally right thing to try to end humanity. What, then, has gone wrong? Before we consider that, let's see how our appeal to risk-weighted expected utility might yield an even stronger and more concerning result.

4 What we together risk

The conclusion of the previous section is a little alarming. According to the methodology favoured by the longtermists, where moral choice is just rational choice with the impersonal utilities demanded by morality, the more 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.[9] 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?

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 credences, but disagree

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

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, since the morally right choice is the one made with the epistemically ideal credences, the morally ideal utilities, and the risk-attitudes 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:[10]

> **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 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.

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

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.

Finally, in many decisions, there is a single population who will be af-

fected 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 A, B, C, and D. 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 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. In Buchak's, that was because the notion of reasonable risk attitudes remained unspecified. In this version, it's because we haven't said 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 A, B, C, and D, 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 longter-

mists 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.

5 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 impersonal good.

(P4) Given your current evidential and historical situation, if you are moderately risk-averse, you maximise your 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 moral utilities, which measure the impersonal good, 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.

5.1 Biting the bullet

Of course, the simplest response is to accept the conclusions. I have nothing to offer as a reply to that. Many people will judge it obviously mistaken, but the effective altruists pride themselves on being unmoved by such intuitions about moral rightness, and this is often a great strength of the movement. And indeed, longtermist thought already runs strongly counter to our intuitive moral judgments, since it asks us to direct our efforts and resources towards a tiny probability of a huge long-term payoff instead of a very high probability of a smaller but still extremely significant short-term payoff, such as saving lives by distributing insecticidal nets to those who need them. What's more, appealing to alternative moral theories to argue for the immorality of bringing humanity to extinction is obviously not going to persuade a utilitarian that this consequence of their moral theory is wrong. So let's leave this and say how we might reject the conclusion.

5.2 Conceptions of the impersonal good

One natural place to look for the argument's weakness is in its axiology. Throughout, we have assumed the austere, monistic conception of impersonal 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 situation A, but also more to contain great suffering in situation B. 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.

Second, we might change what contributes to the impersonal 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 A, 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 A, but it leaves the badness of D unchanged, since the absence of these goods is neutral, and there is nothing that exists in world D that adds further badness to D in the way these higher goods add goodness to A. If these higher goods add enough goodness to A without changing the badness of D, then it may well be that even the risk-averse will prefer the Happy Future Fund over the Quiet End Foundation.

Of course, the most obvious move in this direction is simply to assume that the existence of humanity adds impersonal 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 A, which contains such creatures, but not only would it not add corresponding badness to D; it would in fact add goodness to D, since D contains these beings who boast the special status. And it might add enough goodness to A and D 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 impersonal goodness regardless of how they deploy that intelligence or moral agency. Rather, when we ascribe impersonal value to the existence of humans, we do so because of their potential for doing things that are impersonally 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 D, 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 D 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 impersonal goodness, but rather the degree to which the preferences of the individuals who exist in that situa-

tion 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 A and D, 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 A, since A contains a lot of pleasure and also satisfies the preferences of most people, but it will also boost the badness of D, since D contains a lot of pain and also thwarts the preferences of most people.

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 effective altruist 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 out ending humanity if by doing so you avoid a universally miserable human existence.

5.3 The risk attitudes of possible future people

The stronger conclusion of our argument above is that, not only should a risk-averse individual donate to the Quiet End Foundation instead of the Happy Future Fund, but everyone should, even those who are very strongly risk-inclined. The argument for that relied on two claims. First, a normative claim, namely, clause (iii) of the Risk Principle*. Second, an empirical claim, namely, that the populations in situations A, B, C, and D will contain a substantial proportion of risk-averse individuals whose risk attitudes should lead them to prefer the Quiet End Foundation to the Happy Future Fund. We might support the empirical claim by extrapolating from current populations, all of which seem to contain very substantial proportions of risk-averse individuals. But perhaps there is reason to doubt this projection into the future.

First, you might think that the prevalence of risk aversion in past and current populations is a result of evolutionary pressures exerted by the scarcity of resources, the presence of rare high-stakes decisions, or the pressures of living in small groups (Okasha, 2007; Zhang et al., 2014; Hintze et al., 2015). And you might think that, in scenario A, each of these pressures will no longer exist and there will be enough time in the billion-year future for humans to evolve less risk-averse attitudes. Of course, clause (iii) of the Risk Principle* might still demand an aggregate risk attitude that favours the Quiet End Foundation over the Happy Future Fund. After all, it does say explicitly that, when aggregating the different possible aggregate risk attitudes of the different populations that might be affected, the risk-averse aggregates should receive greater weight than the evidence suggests. And so even if adding many risk-inclined individuals to the population in scenario A gives a risk-inclined aggregate for that outcome, it must still be aggregated with the risk-averse aggregates from the other outcomes, and in a way that gives more weight to those risk-averse aggregates. What's more, the transition away from the evolutionary pressures mentioned above will take some time, and then the evolution of the more risk-inclined attitudes will take further time, and so even if scenario A does include a lot of risk-neutral or risk-inclined individuals in the reasonably far future, it will also continue to include a lot of risk-averse individuals between now and then. And, one further point: it's not clear which evolutionary pressures would bring about this transition to greater risk-inclination; would the lack of the original evolutionary pressures be sufficient?

Nonetheless, let's suppose for a moment that it is plausible that scenario A contains sufficiently many risk-inclined individuals that the Risk Principle* does lead us to aggregated risk attitudes that demand the Happy Future Fund instead of the Quiet End Foundation. Even then, there is another consideration that favours the Quiet End Foundation: the risk-neutral or risk-inclined people whose attitudes skew the aggregate more risk-inclined are the lucky ones; they are the ones inhabiting the good outcome of the decision; they suffer no harm. It is precisely those who will suffer harm who have the risk-averse preferences that would lead them to choose the option that reduces the chance of the scenario they inhabit coming about. It seems that, in such a situation, the risk-averse unlucky individuals should be given even more weight.

5.4 The correct theory of rational choice

The final premise of our argument that we might reject is its endorsement of risk-weighted expected utility theory. We might argue that it is simply not rational to choose in line with its recommendations for risk-averse or risk-inclined individuals. Rationality, we might argue, simply requires you to be risk-neutral and choose in line with standard expected utility theory,

just as the longtermists have always assumed. Indeed, you might take the present argument to be a reductio of Buchak's theory, and add it to other objections that already exist. Buchak has given thorough responses to the main objections to her theory (Buchak, 2013, Chapters 5-7). I will present just one sort of objection and response, which are distinct from but closely related to ones that Buchak considers.

Recall Sheila. She's planning a birding trip and has to choose on Monday between going to nature reserve N1 and nature reserve N2 on Wednesday. And here is the payoff table:

	M	$\overline{M}$
N1	49	49
N2	0	100

But what's more, she's got to choose on Tuesday between going to nature reserve N3 or nature reserve N4 on Thursday. And here is the payoff table for that:

	M	$\overline{M}$
N3	49	49
N4	100	0

That is, at N3, as at N1, Sheila will see 49 new species for sure. But at N4, she'll see 100 species if the winter migration has started and none if it hasn't, the precise reverse of N2. What's more, all the new species she'll see at N3 or N4 are distinct from the species she'll see at N1 or N2. Risk averse as she is, and thinking it just as likely the migration has started as that it hasn't, she will choose N1 on Monday and N3 on Tuesday, thus seeing 98 new species over the course of the two days. But notice: if she'd instead chosen N2 and then N4, as she was perfectly capable of doing, she'd end up seeing 100 species for sure over the course of the two days. That is, Buchak's risk-weighted expected utility theory demands of Sheila, as a risk-averse individual, that she choose a dominated pair of options, N1 then N3; and indeed similar examples can be given for any other sort of risk-sensitive individual, whether risk-averse or risk-inclined, though not for a risk-neutral individual who chooses in line with expected utility theory. Surely this reveals its inadequacy as a decision theory? Or so this objection claims.

My response: how we should respond to this case depends on whether we specify that Sheila knows on Monday that she'll face both the choice between N1 and N2 and then the choice between N3 and N4, or whether we specify that she doesn't know this.

Suppose she doesn't know it. Then it's not clear that there is anything troubling about the fact that Buchak's theory leads her to pick a dominated pair of options. To see this, consider a case in which expected utility theory leads you to pick a dominated pair of options. Suppose your credence

in a proposition changes between one time and another because you gain new evidence between those times and update your credences in the light of that. Then there will be a bet that expected utility theory will tell you to accept at the earlier time, when you have one credence in the proposition, and a different bet it will tell you to accept at the later time, when you have a different credence in that proposition, and yet taken together, this pair of bets will lose you money for sure—that is, accepting both of them is dominated by rejecting them both. But that doesn't show that your change of credence is irrational nor that the decision theory you use to decide whether or not to bet is faulty.

Next suppose that Sheila does know the sequence of choices she'll face: N1 vs N2 on Monday and N3 vs N4 on Tuesday. Then that changes the decision problem she faces on Monday.[11] After all, what Sheila chooses on Monday might affect what she'll choose on Tuesday. And if that's the case, she should choose on Monday knowing that her choice will have that effect on her Tuesday decision, and she should take that into account. For instance, suppose she chooses the safe option N1 on Monday. Then her payoff table on Tuesday is really this:

	M	$\overline{M}$
N1+N3	49+49=98	49+49=98
N1 + N4	49+100=149	49+0=49

Now, since this simply shifts the utility of each option at each state of the world up by the same amount, she will choose N3 (or N1+N3) in this situation. But now suppose she chooses the risky option N2 on Monday. Then her payoff table on Tuesday is really this:

	M	$\overline{M}$
N2 + N3	0 + 49=49	100+49=149
N2 + N4	0+100=100	100+0=100

And, in this case, she will choose N4 (or N2+N4). Knowing all this, on Monday, Sheila's choice is really between N1+N3 and N2+N4, and so her payoff table is really this:

	M	$\overline{M}$
N1 + N3	98	98
N2 + N4	100	100

And so she'll choose N2 on Monday, and then N4 on Tuesday. That is, she'll pick the dominating pair of options. So the threat of the objection recedes.

[11] This observation is sometimes known as *sophisticated choice* in the decision theory literature on choosing when your preferences change, such as in cases of temptation (Hammond, 1988).

6 Conclusion

Thinking about the effects of our actions on the long-term future has led us to an unsettling place. The future might contain vast quantities of great happiness, but it might also contain vast quantities of great misery. Any effort to ensure a long happy future will have to strive both to ensure a long future and to make it happy; and it might succeed at the former without succeeding at the latter. So our efforts to ensure a long happy future will increase the likelihood of such a future, but they will also increase the likelihood of a long miserable future, even if they increase the latter by a smaller amount than the former. And so, if you are risk-averse, or if morality requires you to choose using risk-averse preferences, then you do better to work towards the extinction of humanity. In the final few sections of the paper, I asked whether there might be a problem with the decision theory we use or with the conception of the impersonal good that we feed into it. There is some hope that there is some plausible conception of the impersonal good that assigns sufficiently different moral utilities to the outcomes that, when fed into risk-weighted expected utility theory, require us to donate to the Happy Future Fund instead of the Quiet End Foundation. But they seem to me a little ad hoc and in any case not in the spirit of effective altruism, which is usually more eager to follow where the utilitarian calculus leads than to adjust the axiology to save our intuitions.

We might try to avoid the conclusion by moving away from the consequentialist or decision-theoretic approach to morality that the argument assumes. However, as Greaves & MacAskill (2021) point out, most moral theories would permit other considerations to be overridden when the quantity of impersonal good at stake is of the magnitude considered here. Most people who think that individuals have a right to choose whether or not to try to have a child will think that this right can be overridden in a particular case if for some reason exercising that right would lead to immense misery for a large group of people. So, we might imagine, if the considerations adduced in favour of the Quiet End Foundation are weighty enough, they will override any rights that would be trampled by pursuing that strategy.

So I conclude not satisfied that I have identified any point at which this argument goes wrong, though I remain hopeful that it does. But the existence of this argument, so close in form to the arguments given by the longtermists in favour of devoting resources to those projects rather than to well-understood, near-term health interventions such as distributing insecticidal nets, casts doubt on those arguments.

7 Appendix: the calculations

7.1 The expected utilities of SQ, QEF, and HFF

First, the expected utility of doing nothing (SQ):

$$P(A|\text{SQ})U(A) + P(B|\text{SQ})U(B) + P(C|\text{SQ})U(C) + P(D|\text{SQ})U(D) = pw + qv + ru - st$$

Second, the expected utility of donating to the Quiet End Foundation (QEF):

$$\begin{aligned} P(A|\text{QEF})U(A) + P(B|\text{QEF})U(B) + P(C|\text{QEF})U(C) + P(D|\text{QEF})U(D) = \\ (p-2\varepsilon)w + (q-2\varepsilon)v + (r+6\varepsilon)u + (s-2\varepsilon)(-t) = \\ pw + qv + ru - st - 2\varepsilon w - 2\varepsilon v + 6\varepsilon u + 2\varepsilon t \end{aligned}$$

Third, the expected utility of donating to the Happy Future Fund (HFF):

$$\begin{aligned} P(A|\text{HFF})U(A) + P(B|\text{HFF})U(B) + P(C|\text{HFF})U(C) + P(D|\text{HFF})U(D) = \\ (p+3\varepsilon)w + (q+2\varepsilon)v + (r-6\varepsilon)u + (s+\varepsilon)(-t) = \\ pw + qv + ru - st + 3\varepsilon w + 2\varepsilon v - 6\varepsilon u - \varepsilon t \end{aligned}$$

Now:

(i) QEF $\prec$ SQ iff $-2\varepsilon w - 2\varepsilon v + 6\varepsilon u + 2\varepsilon t < 0$ iff $-w - v + 3u + t < 0$ iff $3u + t < w + v$.

(ii) SQ $\prec$ HFF iff $0 < 3\varepsilon w + 2\varepsilon v - 6\varepsilon u - \varepsilon t$ iff $6u + t < 3w + 2v$.

So QEF $\prec$ SQ $\prec$ HFF iff $3u + t < w + v$ and $6u + t < 3w + 2v$.

7.2 The risk-weighted expected utilities of SQ, QEF, and HFF

First, the risk-weighted expected utility of doing nothing (SQ):

$$\begin{aligned} REU(\text{SQ}) = (1 - R(P(A \vee B \vee C|\text{SQ})))U(\text{SQ \& } D) + \\ (R(P(A \vee B \vee C|\text{SQ})) - R(P(A \vee B|\text{SQ})))U(\text{SQ \& } C) + \\ (R(P(A \vee B|\text{SQ})) - R(P(A|\text{SQ})))U(\text{SQ \& } B) + \\ R(P(A|\text{SQ}))U(\text{SQ \& } A) = \\ (1 - R(p+q+r))(-t) + (R(p+q+r) - R(p+q))u + \\ (R(p+q) - R(p))v + R(p)w \end{aligned}$$

Now suppose $R(x) = x^{3/2}$. Then, now adding the specific numbers we introduced above:

$$\begin{aligned} REU(\text{SQ}) = (1 - (1-10^{-9})^{1.5})(-10^{19}) + \\ ((1-10^{-9})^{1.5} - (1-10^{-9}-10^{-3})^{1.5})10^{6} + \\ ((1-10^{-9}-10^{-3})^{1.5} - (10^{-6})^{1.5})10^{13} + \end{aligned}$$

$$(10^{-6})^{1.5}10^{19} \approx 9,980,003,700,000$$

Second, the risk-weighted expected utility of donating to the Quiet End Foundation:

$$\begin{aligned}REU(\text{QEF}) = {} & (1 - R(P(A \vee B \vee C|\text{QEF})))U(\text{QEF } \& \ D)+ \\ & (R(P(A \vee B \vee C|\text{QEF})) - R(P(A \vee B|\text{QEF})))U(\text{QEF } \& \ C)+ \\ & (R(P(A \vee B|\text{QEF})) - R(P(A|\text{QEF})))U(\text{QEF } \& \ B)+ \\ & R(P(A|\text{QEF}))U(\text{QEF } \& \ A) = \\ & (1 - R(p + q + r + 2\varepsilon))(-t)+ \\ & (R(p + q + r + 2\varepsilon) - R(p + q - 4\varepsilon))u+ \\ & (R(p + q - 4\varepsilon) - R(p - 2\varepsilon))v + R(p - 2\varepsilon)w\end{aligned}$$

Now suppose $R(x) = x^{\frac{3}{2}}$. Then, now adding the specific numbers we introduced above:

$$\begin{aligned}&REU(\text{QEF}) = \\ &\quad (1 - (1 - 10^{-9} + (2 \times 10^{-10}))^{1.5})(-10^{19})+ \\ &\quad\quad ((1 - 10^{-9} + (2 \times 10^{-10}))^{1.5} - (1 - 10^{-9} - 10^{-3} - (4 \times 10^{-10}))^{1.5})10^{6}+ \\ &\quad\quad\quad ((1 - 10^{-9} - 10^{-3} - (4 \times 10^{-10}))^{1.5} - (10^{-6} - (2 \times 10^{-10}))^{1.5})10^{13}+ \\ &\quad\quad\quad\quad (10^{-6} - (2 \times 10^{-10}))^{1.5}10^{19} \approx 9,983,007,000,000\end{aligned}$$

Third, the risk-weighted expected utility of donating to the Happy Future Fund:

$$\begin{aligned}REU(\text{HFF}) = {} & (1 - R(P(A \vee B \vee C|\text{HFF})))U(\text{HFF } \& \ D)+ \\ & (R(P(A \vee B \vee C|\text{HFF})) - R(P(A \vee B|\text{HFF})))U(\text{HFF } \& \ C)+ \\ & (R(P(A \vee B|\text{HFF})) - R(P(A|\text{HFF})))U(\text{HFF } \& \ B)+ \\ & R(P(A|\text{HFF}))U(\text{HFF } \& \ A) = \\ & (1 - R(p + q + r - \varepsilon))(-t)+ \\ & (R(p + q + r - \varepsilon) - R(p + q + 5\varepsilon))u+ \\ & (R(p + q + 5\varepsilon) - R(p + 3\varepsilon))v + R(p + 3\varepsilon)w\end{aligned}$$

Now suppose $R(x) = x^{\frac{3}{2}}$. Then, now adding the specific numbers we introduced above:

$$\begin{aligned}&REU(\text{HFF}) = \\ &\quad (1 - (1 - 10^{-9} - 10^{-10})^{1.5})(-10^{19})+ \\ &\quad\quad ((1 - 10^{-9} - 10^{-10})^{1.5} - (1 - 10^{-9} - 10^{-3} + (5 \times 10^{-10}))^{1.5})10^{6}+ \\ &\quad\quad\quad ((1 - 10^{-9} - 10^{-3} + (5 \times 10^{-10}))^{1.5} - (10^{-6} + (3 \times 10^{-10}))^{1.5})10^{13}+ \\ &\quad\quad\quad\quad (10^{-6} + (3 \times 10^{-10}))^{1.5}10^{19} \approx 9,978,508,200,000\end{aligned}$$

So:

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

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.

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

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.

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

Hammond, P. (1988). Consequentialist Foundations for Expected Utility. *Theory and Decision*, *25*, 25–78.

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

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.

Sharples, F., Husbands, J., Mazza, A.-M., Thevenon, A., & Hook-Barnard, I. (2015). Potential Risks and Benefits of Gain-of-Function Research: Summary of a Workshop. Tech. rep., National Research Council.

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.