Chapter 6

Tiny Probabilities and the Value of the Far Future[*]

ABSTRACT: Morally speaking, what matters the most is the far future—at least according to Longtermism. The reason why the far future is of utmost importance is that our acts' expected influence on the value of the world is mainly determined by their consequences in the far future. The case for Longtermism is straightforward: Given the enormous number of people who might exist in the far future, even a tiny probability of affecting how the far future goes outweighs the importance of our acts' consequences in the near term. However, it seems that there is something wrong with a theory that lets very small probabilities of huge payoffs dictate one's course of action. If, instead, we discount very small probabilities down to zero, we may have a response to Longtermism provided that its truth depends on tiny probabilities of vast value. Contrary to this, I will argue that discounting small probabilities does not undermine Longtermism.

[*] I wish to thank Gustav Alexandrie, Andreas Mogensen, Teruji Thomas, Hayden Wilkinson, participants of GPI's Early Career Conference Programme 2021 and the audience of the 8th Oxford Workshop on Global Priorities Research for valuable feedback.

Morally speaking, what matters the most is the far future—at least according to the following view:[1]

> **Longtermism:** In the most important decision situations, our acts' expected influence on the value of the world is mainly determined by their possible consequences in the far future.

On this view, the far future is of utmost importance. In the most important decision situations, we can often simply ignore our acts' effects in the near future and instead focus on their effects in the distant future. Longtermism follows naturally from additive views of value, such as total utilitarianism. Given the enormous number of people who might exist in the far future, even a tiny probability of affecting how the far future goes outweighs the importance of our acts' consequences in the near term.[2] So, if we are in a position to foreseeably affect the far future, our influence in the near term is outstripped by our influence in the far future.[3] However, one might reasonably doubt that we can have probabilistic evidence for some acts resulting in better outcomes than the alternatives hundreds or thousands of years from now.

One way in which we might beneficially influence the far future—it has been

[1] MacAskill (2019) and Greaves and MacAskill (2021). Greaves and MacAskill (2021, p. 2) define Longtermism as the view according to which we should be particularly concerned with ensuring that the far future goes well, and *Strong Longtermism* as the view on which the impact on the far future is the most important feature of our actions today. They defend axiological and deontic versions of this thesis. The former states that far-future effects are the most important determinant of the value of our options, while the latter states that they are the most important determinant of what we ought to do. See Greaves and MacAskill (2021, p. 3). For discussions of related topics, see for example Bostrom (2003), Beckstead (2013) and Ord (2020).

[2] Greaves and MacAskill (2021, p. 1).

[3] 'Foreseeably' in this context means probabilistic evidence rather than certainty or knowledge. We need not be able to foresee the effects of our actions so long as we can assign probabilities to the possible outcomes, conditional on the available acts, such that the expected value is favorable.

argued—is by mitigating existential risks.[4] Existential risks are risks that threaten the destruction of humanity's long-term potential. Such risks might be posed by, for example, synthetic pathogens, artificial intelligence (AI) systems, asteroids or climate change. Extinction risks are one type of existential risk. Because humanity's future is potentially very long, even relatively small reductions in the net probability of existential catastrophe correspond to enormous increases in expected moral value.[5] So, it can be argued that even very small reductions of existential risk have an expected moral value greater than that of the provision of any near-term good, such as the direct benefit of saving one billion present-day lives.[6]

However, it seems that there is something wrong with a theory that lets tiny probabilities of huge value dictate one's course of action. At least, such a theory would give counterintuitive recommendations. Consider, for example, the following case:[7]

> **Pascal's Mugging:** A stranger approaches Pascal and claims to be an Operator from the Seventh Dimension. He promises to perform magic that will give Pascal an extra thousand quadrillion happy days in the Seventh Dimension if Pascal pays the mugger ten livres—money

[4] Bostrom (2013). I will focus on existential risk mitigation as it seems one of the best candidates for longtermist interventions in terms of importance and tractability. Some longtermists focus instead on positively influencing humanity's trajectory conditional on survival.

[5] Bostrom (2013).

[6] Bostrom (2013, pp. 18–19). Greaves and MacAskill (2021, p. 11) write that even if there are 10^{14} lives to come (one of their more conservative estimates), a one-millionth of one percentage point reduction in the near-term extinction risk would be equivalent to the value of a million lives saved. On their main estimate of 10^{24} expected future lives, this becomes 10^{16} lives saved.

[7] Bostrom (2009). This case is based on informal discussions by various people, including Eliezer Yudkowsky (2007).

> that the mugger will use for helping very many orphans in the Seventh Dimension.

Pascal thinks that the stranger is almost certainly lying. However, the possible payoff is so enormous that he is forced to conclude that the expected utility of paying the mugger is positive.[8] Importantly, the mugger points out that as long as Pascal gives a non-zero probability to the mugger being able to reward him with any finite amount of utility, the mugger can increase the payoff until the offer has positive expected utility.[9] Consequently, expected utility maximization (with no bound on utilities) recommends that Pascal pay the mugger—and thus, it gives the intuitively wrong recommendation.

Another version of this case is relevant to the topic of this chapter. In this case, the mugger exploits Pascal's expected-utility-maximizing descendant by utilizing research on existential risk and the long-term potential of humanity:[10]

> **Pascal's Mugger Strikes Again:** A stranger in a pub tells Pascal that a secretive organization is preparing a deadly disease that will make Earth uninhabitable within the next two years. However, the brewery that makes a particular ale sold at the pub also develops cutting-edge vaccines, and they need another £2 to pay for their electricity bills, or else their supplier will shut the factory off. The stranger forgot his

[8] 'Utility' here can be interpreted as moral value or as a decision-theoretic construct representing the betterness of prospects. Moral value, in turn, should be understood as reflecting the importance or significance of an act or outcome from a moral perspective.

[9] This may not be true if utility is bounded as standard axiomatizations of expected utility maximization require. See for example Kreps (1988, p. 63) and §1 and §2.1 in Chapter 1 of this thesis.

[10] This case is from Balfour (2021).

> wallet at home but—he claims—Pascal can save humanity from this deadly disease by buying him a pint of this ale.

Again, Pascal thinks that the mugger is almost certainly lying. However, given that the future of humanity is at stake, buying a pint might be the right course of action.[11] The mugger also warns that Pascal will be mugged every waking moment for the rest of his life, not by the mugger, but by the future of humanity itself. The mugger argues that, as an expected utility maximizer, Pascal must always perform the action which seems least likely to condemn humanity to extinction: "[Y]ou'll need to maintain constant vigilance, thinking constantly about which of your actions is least likely to destroy humanity."[12]

These cases are silly. If one were confronted with claims such as these muggers', one would consider them outlandish. However, there are reasons to think that even outlandish propositions should be assigned a non-zero probability. For example, according to Bayesianism, conditionalization is the right way to respond to new evidence. So, on this view, if one assigns some proposition zero (subjective) probability, one will always continue to do so no matter the evidence one obtains. However, sufficiently strong evidence should convince one of the truths of even outlandish propositions. If the mugger takes Pascal for a visit in the Seventh Dimension, Pascal should consider the mugger's original offer more probable than before, and in

[11] One could object that, instead of buying a pint for the stranger, Pascal should donate that money to some organization that works to mitigate existential risk, as this is a more effective way of securing humanity's future. That seems right. However, if Pascal knows that he will not do so, then actualism would advise Pascal to buy a pint for the stranger.

[12] Balfour (2021, p. 123).

particular, not consider it impossible.[13] Therefore, even outlandish propositions should be assigned non-zero probabilities, albeit tiny ones.[14] However, provided that the probabilities and the utilities work out the right way, expected utility maximization (with no bound on utilities) implies that Pascal should pay the mugger.[15] More generally, it leads to

> **Probability Fanaticism:** For any tiny probability $p > 0$, and for any finite utility u, there is some large enough utility U such that probability p of U (and otherwise nothing) is better than certainty of u.[16]

In response to cases that involve tiny probabilities of huge payoffs, some have argued that we ought to discount very small probabilities down to zero—let's call this *Probability Discounting*.[17] If we are indeed rationally required or permitted to discount small probabilities, then we may have an argument against Longtermism provided that its truth depends on tiny probabilities of huge value. In fact, this may

[13] Pascal might still think that he was probably, for example, hallucinating rather than visiting the Seventh Dimension. However, if the mugger gave him the ability to visit the Seventh Dimension repeatedly, he should not consider the mugger's original proposition impossible, even if hallucinating is still the most likely explanation.

[14] For a related discussion, see Francis and Kosonen (n.d.).

[15] Why not just bound utilities? This seems implausible, at least when it comes to ethical decisions. For example, this theory would imply that it is better to save some (very large) number n of lives for sure than to save *any number* of lives with a probability of almost one. See §4 in the introduction of this thesis.

[16] Wilkinson (2022, p.449). For discussions related to Probability Fanaticism, see Beckstead (2013, ch. 6), Beckstead and Thomas (2020), Goodsell (2021), Russell and Isaacs (2021) and Russell (2021).

[17] Monton (2019) argues that very small probabilities should be discounted down to zero, while Smith (2014) argues that one is rationally permitted—but not required—to do so. Smith argues that discounting very small probabilities allows one to get a reasonable expected utility for the Pasadena game (see [Nover and Hájek 2004]). See Hájek (2014), Isaacs (2016) and Lundgren and Stefánsson (2020) for criticisms of discounting small probabilities.

be one of the most plausible ways in which the argument for Longtermism might fail.[18] As mentioned above, one possible longtermist cause area is the mitigation of existential risk. However, the actions of a single individual are very unlikely to affect whether an existential catastrophe occurs.[19] The argument for prioritizing such actions is that if they make a difference, they might make an enormous one, such as delay human extinction by centuries, millennia, or more.[20]

This chapter argues that Probability Discounting does not undermine Longtermism. Even if one accepts a view on which small probabilities should be discounted down to zero, one should still consider the far future to be of utmost importance (or reject Longtermism for some other reason). I will discuss three arguments against Longtermism from discounting small probabilities. §2 discusses the argument that the probabilities of existential catastrophes are so low that one ought to ignore them. §3 discusses the argument that once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of lives in the far future is too small for Longtermism to be true. Lastly, §4 and §5 discuss the argument that the probability that an agent makes a difference to whether an existential catastrophe occurs or not is so small that it should be ignored. This chapter concludes that none of these arguments undermine Longtermism. Before going into these arguments, I will first say more about Probability Discounting. This

[18] Greaves and MacAskill (2021, p. 25). Besides discounting small probabilities, one could avoid letting tiny probabilities of huge value dictate one's course of action by having a bounded utility function. See for example Beckstead and Thomas (2020).

[19] In contrast, for some suitably capacious 'we,' we together might be likely to make a difference to net existential risk. I will discuss this in §4 and §5 of this Chapter.

[20] Greaves and MacAskill (2021).

chapter focuses on three versions of Probability Discounting: Naive Discounting, Tail Discounting and State Discounting.[21] Next, I will introduce Naive Discounting.

1 Discounting small probabilities

This section introduces one of the simplest versions of Probability Discounting. It also discusses choosing the threshold below which probabilities are small enough to be ignored.

Probability Discounting was originally proposed by Nicolaus Bernoulli.[22] He writes: "[T]he cases which have a very small probability must be neglected and counted for nulls, although they can give a very great expectation."[23] But when are probabilities small enough to be discounted? Or, as Buffon writes, "one can feel that it is a certain number of probabilities that equals the moral certainty, but what number is it?"[24] Some have suggested possible discounting thresholds. For Buffon and Condorcet, the discounting thresholds were 1 in 10,000 and 1 in 144,768 (respectively). Buffon chose his threshold because it is the probability of a 56-year-

[21] For a discussion of the different versions of Probability Discounting, see Chapter 4 of this thesis.

[22] Monton (2019) calls discounting small probabilities 'Nicolausian discounting' after Nicolaus Bernoulli.

[23] Pulskamp (n.d., p. 2). Discounting small probabilities is Bernoulli's solution to the St. Petersburg paradox.

[24] Hey et al. (2010, p. 256). Nicolaus Bernoulli also raised this problem: "It is necessary [...] to determine as far as where the quantity of a probability must diminish, so that it be able to be counted null." See Pulskamp (n.d., p. 5).

old man dying in one day—an outcome reasonable people usually ignore.[25] Condorcet's justification for his threshold is that 1 in 144,768 is the difference between the probability that a 47-year-old man would die within 24 hours and the probability that a 37-year-old man would, and that difference would not keep anyone awake at night.[26]

It seems implausible that agents are rationally required to use some particular discounting threshold. Monton, who defends Probability Discounting, agrees. He argues that the discounting threshold is subjective within reason.[27] He would consider a threshold of $1/2$ irrational and some astronomically small threshold unreasonable. Nevertheless, there is no particular discounting threshold that all agents are rationally required to use. For Monton, the discounting threshold is approximately 1 in 2 quadrillion.[28] His justification for this threshold is that 1 in 2 quadrillion is between $1/2^{50}$ and $1/2^{51}$, and he treats the probability of getting tails at least 50 times in a row (with a fair coin) as a probability-zero event.

So, Probability Discounting is the idea that one should ignore sufficiently small probabilities—but small probabilities of *what*? On one version of this view, we should ignore *outcomes* associated with tiny probabilities. There is some threshold probability t such that outcomes whose probabilities are below this threshold are ignored.[29] Ignoring such outcomes can be done by conditionalizing on

[25] Hey et al. (2010, p. 257). See Monton (2019, pp. 8–9) for a discussion of Buffon's view.

[26] See Monton (2019, pp. 16–17).

[27] Monton (2019, §6.1) Note that this threshold may also be vague. See Lundgren and Stefánsson (2020, p. 911).

[28] Monton (2019, p. 17).

[29] Alternatively, one might have a threshold probability t such that outcomes whose probabilities

the supposition that an outcome of non-negligible probability occurs, where an 'outcome of non-negligible probability' is one whose associated probability is at least as great as the discounting threshold.[30] After conditionalization, options are compared by their 'probability-discounted expected utilities'. Let $X \succsim Y$ mean that X is at least as preferred as Y, and let $EU(X)_{pd}$ mean the expected utility of prospect X when tiny probabilities have been discounted down to zero (read as 'the probability-discounted expected utility of X'). Then, this version of Probability Discounting—let's call it *Naive Discounting*—states the following:[31]

> **Naive Discounting:** First, conditionalize on obtaining some outcome of non-negligible probability. Then, for all prospects X and Y, $X \succsim Y$ if and only if $EU(X)_{pd} \geq EU(Y)_{pd}$.

To summarize, Probability Discounting is the idea that very small probabilities should be ignored in practical decision-making. One of the simplest versions of this view is Naive Discounting, on which one should conditionalize on not obtaining outcomes associated with negligible probabilities. Next, I will consider an argument against Longtermism that someone with this view might give.

are at most as great as this threshold are ignored, but outcomes whose probabilities are greater than the threshold are not ignored.

[30] Smith (2014, p. 478).

[31] See §1 in Chapter 4 of this thesis.

2 Probability of an existential catastrophe

This section discusses the argument that the probabilities of existential catastrophes are so low that we should ignore them. However, it seems that even in the next century, existential risks have probabilities that are above any reasonable discounting thresholds. Naive Discounting faces a problem with individuating outcomes, so it is unclear what it says. Naive Discounting also violates dominance. Tail Discounting is a more plausible view, as it solves the outcome individuation problem and does not violate dominance. However, Tail Discounting does not ignore near-term extinction risks, so it does not undermine Longtermism in this way.

2.1 Existential risks in this century

It might be argued that existential catastrophes are so unlikely that we should ignore them—let's call this the *Low Risks Argument*.

> **Low Risks Argument:** The probabilities of existential risks are so tiny that we should ignore existential risks; we should evaluate options as though those risks are guaranteed not to eventuate.

This argument requires a reference to some time period: What is the relevant time period during which existential risks are unlikely to occur? After all, eventually, humanity will (almost certainly) go extinct. However, even in the next century, the net existential risk seems non-negligible. Ord (2020, p. 167) estimates that the probability of an existential catastrophe within the next 100 years is 1/6—way above any reasonable discounting threshold. The British Astronomer Royal Sir Martin

Rees has an even more pessimistic view. Rees (2003, p. 8) writes: "I think the odds are no better than fifty-fifty that our present civilization on Earth will survive to the end of the present century." Ord (2020, p. 167) gives the following estimates for existential catastrophes from specific causes within the next 100 years: 1 in 1,000,000 from asteroid or comet impact, 1 in 30 from engineered pandemics and 1 in 10 from unaligned AI (see table 1). Other estimates for *extinction* risks in the next 100 years are, for example, 1 in 15 billion from a 10 km+ asteroid colliding with the Earth,[32] between 1 in 600,000 and 1 in 50 from an extinction-level pandemic,[33] and a very conservative assessment would assign at least a 1 in 1000 chance to an AI-driven catastrophe that is as bad or worse than human extinction.[34]

Table 1
Existential and Extinction Risks in the Next 100 Years

	Existential risk (Ord, 2020)	Extinction risk (Others)
Asteroids	1 in 1,000,000*	1 in 15 billion
Pandemics	1 in 30**	1 in 600,000 to 1 in 50
AI	1 in 10	$\geq$ 1 in 1000

*=including comets, **=engineered pandemics.

If we individuate outcomes as 'human extinction from an asteroid impact in the

[32] The risk of a 10 km+ asteroid colliding with the Earth is estimated to be 1 in 150 million. See Ord (2020, p. 71). It is estimated that an asteroid with a 10 km+ diameter has at least a 1% chance of causing human extinction. See Newberry (2021, p. 3).

[33] Millett and Snyder-Beattie (2017).

[34] Greaves and MacAskill (2021, pp. 14–15). The expert median estimate for an AI-driven catastrophe is 5%. See Grace et al. (2018, p. 733).

next 100 years,' 'extinction-level pandemic in the next 100 years' and so on, then some extinction (and existential) risks are plausibly non-negligible. One should not ignore, for example, a 1 in 1000 chance of an AI-driven catastrophe in the next 100 years. However, if we individuate outcomes as 'extinction due an asteroid impact on the 4th of January 2055 at 13:00–14:00', 'extinction due to an asteroid impact on the 4th of January 2055 at 14:00–15:00' and so on, then extinction (and existential) risks might be negligible. It is difficult to see what the privileged way of individuating outcomes would be, and choosing one way over the others seems arbitrary. More generally, Naive Discounting faces the following problem:[35]

> **Outcome Individuation Problem:** If we individuate outcomes with too much detail, all outcomes have negligible probabilities. Is there a privileged way of individuating outcomes that avoids this?

If there is a plausible solution to the Outcome Individuation Problem, this solution should not tell one to ignore a net existential risk of 1/6 or a 1/10 risk of an AI-driven catastrophe.[36] Consequently, Naive Discounting does not undermine Longtermism, at least in this way. However, these relatively high estimates of existential risks have also been questioned.[37] Might we, after all, have a challenge to

[35] See also Beckstead and Thomas (2020, p. 13).

[36] One possible solution is to individuate outcomes by their utilities. See §1 in Chapter 4 of this thesis. However, this solution would imply that a human extinction on the 15th of February 2022 and one on the 16th February 2022 are distinct outcomes, given that their values are slightly different. Consequently, all possible extinction outcomes might have negligible probabilities, even if net extinction risk is high. This would secure the result that Probability Discounting undermines Longtermism. However, first individuating outcomes in this way and then applying Probability Discounting is absurd because net extinction risk could be arbitrarily high.

[37] See Luisa Rodriguez (2021) on the 80,000 Hours podcast for an informal discussion on this

Longtermism?

2.2 Tail Discounting

In addition to the Outcome Individuation Problem, Naive Discounting also faces other problems. For example, it violates dominance.[38] Instead, one might accept *Tail Discounting*, which states that one ought to ignore both the left and the right 'tails' of the distribution of possible outcomes when these outcomes are ordered by one's preference.[39] Tail Discounting is a more plausible version of Probability Discounting than Naive Discounting. However, it does not undermine Longtermism, even if the probability of an existential catastrophe is tiny.

Call the outcomes that fall in the middle of the distribution of possible outcomes 'normal outcomes'. Then, Tail Discounting states the following:[40]

topic. Rodriguez argues that humanity has a high probability of recovery from a non-extinction catastrophe and that for many of the threats, it is difficult to imagine a single sudden cataclysm that kills literally everyone.

[38] For example, Naive Discounting judges a prospect that saves a life with a negligible probability (and otherwise nothing happens) as equally good as a prospect that certainly saves no one. Using very-small-probability outcomes as tiebreakers (as 'Lexical Discounting' does) still violates Statewise Dominance in a more complicated case. See §2 in Chapter 4 of this thesis. Also see Isaacs (2016), Smith (2016), Monton (2019, pp. 20–21), Lundgren and Stefánsson (2020, pp. 912–914) and Beckstead and Thomas (2020, §2.3) on discounting small probabilities and dominance violations.

[39] Beckstead and Thomas (2020, 2.3). Unless one considers very-small-probability outcomes in cases of ties (as the definition of Tail Discounting given in this chapter does), Tail Discounting violates dominance reasoning.

[40] More formally, this view states the following:

> **Tail Discounting:** To determine $EU(X)_{pd}$, first order the possible outcomes of some prospect X from the least to the most preferred. Then, conditionalize on obtaining some outcome in the middle part of the distribution such that the following necessary conditions hold for all outcomes o that are not ignored:
>
> i The probability of obtaining an outcome that is at least as good as o is above the discounting threshold and

Tail Discounting: For all prospects X and Y, $X \succsim Y$ if and only if

- $EU(X)_{pd} > EU(Y)_{pd}$ or
- $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,

where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that a normal outcome occurs.

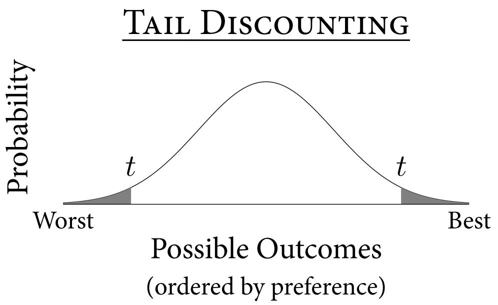
Tail Discounting solves the Outcome Individuation Problem because, on this view, it does not matter how finely outcomes are individuated; one always ignores the tails of the distribution of possible final values. When the possible outcomes of a prospect are ordered by one's preference, the order of these outcomes will not change by individuating these outcomes more finely.

Next, suppose the possible outcomes of some prospect are normally distributed when they are ordered from the least to the most preferred. Then, Tail Discounting tells us to ignore the grey areas under the curve (the discounting threshold is denoted by t):

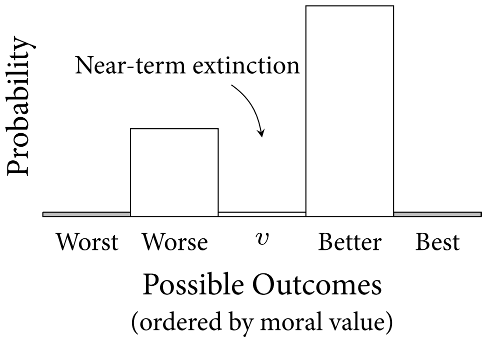
ii the probability of obtaining an outcome that is at most as good as o is above the discounting threshold.

If some outcome o fulfills the above necessary conditions, and

- the probability of obtaining an outcome that is better than o is below the discounting threshold, then decrease the probability of obtaining o until the total discounted probability of outcomes that are at least as good as o equals the discounting threshold (and conditionalize to make sure the remaining probabilities add up to 1), and
- if the probability of obtaining an outcome that is worse than o is below the discounting threshold, then decrease the probability of obtaining o until the total discounted probability of outcomes that are at most as good as o equals the discounting threshold (and conditionalize to make sure the remaining probabilities add up to 1).



What does Tail Discounting say about extinction risks? Suppose that the moral value of a near-term extinction is v. As long as v falls in the middle of the distribution of possible outcomes' values, Tail Discounting will not ignore the possibility of a near-term extinction. If there are non-negligible probabilities of worse and better outcomes than a near-term extinction, then near-term extinction scenarios may fall somewhere in the middle of the distribution of possible outcomes. Consider for example the following prospect:



In this case, the probability of a near-term extinction is tiny. However, the probability of obtaining an outcome that is at least as good as a near-term extinction is above the discounting threshold. Similarly, the probability of obtaining an outcome that is at most as good as a near-term extinction is also above the discounting threshold. Consequently, Tail Discounting recommends against ignoring the possibility of a near-term extinction. Even if there is just a small probability of a near-term extinction and one can decrease this probability by just a small amount, Tail Discounting advises one to mitigate this risk (as long as the probabilities and the utilities work out the right way).

It seems plausible that the probabilities of both better and worse futures than a near-term extinction are above reasonable discounting thresholds. For example, the value of the world might be negative due to human and non-human animal suffering and continue to be negative in the future. Thus, there is a non-negligible probability that the future is worse than a near-term extinction. On the other hand, the value of the world might be net positive and continue to be so in the future. Alternatively, technological progress might increase well-being and create an overall positive future. Thus, there is a non-negligible probability that the future is better than a near-term extinction. Both better and worse possibilities seem non-negligible; neither is very unlikely. Consequently, someone who accepts Tail Discounting will not ignore the possibility of a near-term extinction. Tail Discounting only ignores outcomes with extreme values, and a near-term extinction event—plausibly—is not one.[41]

[41] One might object: So much the worse for Tail Discounting! By not advising one to ignore very-small-probability outcomes, such as (possibly) a human extinction, it fails to adequately capture the intuition behind Probability Discounting. Instead, one might accept a version of Tail Discounting on which one compares every prospect to some baseline prospect in the following way: First, calculate the differences in utilities between a given prospect and the baseline prospect in each state of nature. Next, order these differences from the largest loss to the largest gain. Then, ignore the right and left tails of this distribution. In effect, one is ignoring the possibility of a given prospect chang-

To summarize, I have discussed the Low Risks Argument: The probabilities of existential catastrophes are so low that we ought to ignore them. However, it seems that, even in the next century, the net existential risk and some specific existential risks have probabilities above any reasonable discounting thresholds. Naive Discounting faces the Outcome Individuation Problem, so it is unclear what it says; one can individuate existential catastrophes arbitrarily finely, and depending on how they are individuated, their associated probabilities may fall above or below the discounting threshold. However, an acceptable solution to this problem should not imply that one ought to ignore a net existential risk of 1/6 in the next century. Tail Discounting is more plausible than Naive Discounting, as it solves the Outcome Individuation Problem and does not violate dominance. However, as long as there are non-negligible probabilities of better and worse outcomes than a near-term extinction, Tail Discounting will not ignore near-term extinction risks, even if their associated probabilities are negligible.

To conclude, the Low Risks Argument does not undermine Longtermism. The next section discusses a second argument against Longtermism.

ing the value of the world by much. A prospect that lowers the probability of a near-term extinction will have a much higher value than the baseline prospect in some state of nature (namely, the state in which an extinction would have happened had the agent done nothing). This view—called *Baseline Tail Discounting*—will then ignore this large difference in value, assuming that it falls in the tail of the distribution of value differences. See Chapter 4 of this thesis on Baseline Tail Discounting (and a related view called Baseline Stochastic Discounting). However, the argument in §5 of the current chapter also shows (changing what needs to be changed) that Baseline Tail Discounting does not undermine Longtermism.

3 Size of the future

This section discusses the argument that once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of individuals in the far future is too small for Longtermism to be true. Contrary to this, I will argue that there are enough individuals in the far future in expectation for Longtermism to be true even if one accepts Probability Discounting.

3.1 Expected population sizes required for Longtermism

For Longtermism to hold, it also needs to be true that there is in expectation a sufficient number of individuals in the far future.[42] If in expectation the number of individuals is small no matter what we do, then it will not be true that even relatively small changes in the probability of an existential risk have great expected value. So, the argument goes, once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of future people becomes too small—let's call this the *Small Future Argument*.

> **Small Future Argument:** Once we ignore unlikely scenarios, the expected number of individuals in the far future is too small for Longtermism to be true.

Next, I will discuss whether or not there are enough individuals in the far future for existential risk mitigation to have a higher expected value than the neartermist

[42] More precisely, it is not the number of individuals but the amount of value that matters. There might be a great quantity of value in the far future even if the number of individuals is relatively small if these individuals live very long lives. See for example Gustafsson and Kosonen (n.d.).

causes. The cost-effectiveness of antimalarial bednet distribution may be used as an upper bound to attainable near-term benefits per unit of spending.[43] The distribution of insecticide-treated bednets in malarial regions saves a life on average for a little over $4000.[44] Suppose Shivani is thinking how to improve the world the most with her $10,000.[45] By donating to the Against Malaria Foundation, she can save in expectation 2.5 lives. Suppose that Longtermism is true in Shivani's situation if and only if, in expectation, more than 2.5 additional lives exist in the far future if she donates to some longtermist cause.[46]

An example of existential risk mitigation that longtermists might focus on is the detection and potential deflection of asteroids.[47] It is estimated that NASA's Spaceguard Survey, which tracks near-Earth objects in order to identify any on impact trajectories, reduced extinction risk by at least 1 in 2000 trillion per $100 spent.[48,49]

[43] Greaves and MacAskill (2021, p. 2).

[44] GiveWell (2020).

[45] This case is modified from Greaves and MacAskill (2021).

[46] Note that longtermist causes typically also create near-term benefits, and these near-term benefits might be great enough for existential risk mitigation to pass a cost-effectiveness analysis even if one ignores the far future effects of one's acts. Moreover, even if the near-term benefits are not sufficient on their own, the far future effects might add just enough expected value to make existential risk mitigation the best course of action (even though most of the expected influence of existential risk mitigation comes from near-term effects). This is irrelevant to whether or not one should mitigate existential risks, but it matters to whether Longtermism is true. This point is important. Even if Longtermism turns out to be false, existential risk mitigation might still be the right course of action. It is also worth noting that paradigmatic neartermist causes, such as distributing anti-malarial bednets, can also have foreseeable long-term effects.

[47] Greaves and MacAskill (2021, p. 11).

[48] Greaves and MacAskill (2021, p. 11). Coincidentally, this is Monton's threshold for discounting small probabilities ($5 \cdot 10^{-16}$). See Monton (2019, p. 17).

[49] Interestingly, in 2022 NASA will redirect an asteroid for the first time in human history (by slamming a spacecraft into it) for testing technologies that we may need in the future. Their target is a 500-foot-wide moon orbiting a half-mile-wide asteroid called Didymos. This moon is roughly the size of an asteroid that can obliterate cities. See Drake (2020).

But further work on asteroids is expected to have lower cost-effectiveness.[50] It is estimated that a 10 km+ asteroid has at least a 1% chance of causing human extinction if it collides with the Earth.[51] While the probability of a 10 km+ asteroid colliding with the Earth is on average 1 in 1.5 million per century, astronomers are confident that they have found all 10 km+ asteroids in at least 99% of the sky.[52] The remaining risk of a 10 km+ asteroid colliding with the Earth in the next 100 years is estimated to be 1 in 150 million.[53] Consequently, the probability of human extinction from an asteroid impact in the next 100 years is 1 in 15 billion.

The cost of detecting (with almost certainty) any remaining 10 km+ asteroids is estimated to be at most $1.2 billion, and we might assume that we can reduce extinction by 5% (relatively) if we detect one on a collision course.[54] Shivani's proportion of the $1.2 billion required to reduce the risk to (near) zero is 1/120,000. It is plausible that she would reduce the risk by the same proportion, that is, by 1 in 2.4 million.[55] Consequently, by donating $10,000 to asteroid detection, Shivani can provide a 1 in 33,000 trillion absolute reduction in the probability of extinction from an asteroid collision in the next 100 years.[56]

Another possible longtermist cause area is the prevention of extinction-level

[50] Greaves and MacAskill (2021, p. 11).

[51] Newberry (2021, p. 3).

[52] Ord (2020, p. 71).

[53] Ord (2020, p. 71).

[54] Newberry (2021, pp. 5–6). Inspired by the movie Don't Look Up, Lubin and Cohen (n.d.) estimate that humanity could, in theory, defend itself against a comet of a 10km diameter using existing technology even in the extreme case where it is detected just six months before impact.

[55] Greaves and MacAskill (2021, p. 16). $0.05 \cdot 10000/(1.2 \cdot 10^9) \approx 4 \cdot 10^{-7}$.

[56] $1/(15 \cdot 10^9) \cdot 0.05 \cdot 10000/(1.2 \cdot 10^9) \approx 3 \cdot 10^{-17}$ (1 in 33,000 trillion).

pandemics.[57] The risk of an extinction-level pandemic in the next 100 years is estimated to be between 1 in 600,000 and 1 in 50.[58] Taking the geometric mean of the two methods that generate the lower estimates for extinction risk gives a probability of about 1 in 22,000 for extinction from a pandemic over the next 100 years.[59] It is estimated that $250 billion spent on strengthening healthcare systems would reduce the chance of an extinction-level pandemic in the next 100 years by at least a proportional 1%.[60] Consequently, by donating $10,000 to pandemic prevention, Shivani can provide a 1 in 2.5 billion relative reduction and a 1 in 50 trillion absolute reduction in the probability of an extinction-level pandemic in the next 100 years.[61]

Lastly, another possible longtermist cause area is the prevention of an existential catastrophe due to artificial general intelligence.[62] In the most comprehensive study of its kind, AI experts estimated that the probability of an extremely bad outcome, such as human extinction, due to high-level machine intelligence (at any point in time) is 5%.[63] The same experts gave a 50% chance for high-level machine

[57] Greaves and MacAskill (2021, p. 12).

[58] Millett and Snyder-Beattie (2017).

[59] Greaves and MacAskill (2021, p. 12).

[60] Millett and Snyder-Beattie (2017, p. 379).

[61] $0.01 \cdot 10000/(250 \cdot 10^9) \approx 4 \cdot 10^{-10}$ (1 in 2.5 billion). $1/22000 \cdot 0.01 \cdot 10000/(250 \cdot 10^9) \approx 2 \cdot 10^{-14}$ (1 in 50 trillion).

[62] See for example Greaves and MacAskill (2021, pp. 14–15). GPT-3 (n.d.) disagrees: “There is no evidence that artificial general intelligence (AGI) is an existential threat. AGI has the potential to cause a lot of harm, but so far there is no evidence that it will be able to achieve a level of intelligence that would allow it to cause existential harm.”

[63] Grace et al. (2018, p. 733). “High-level machine intelligence” is achieved when unaided machines can accomplish every task better and more cheaply than human workers. See Grace et al. (2018, p. 731).

intelligence occurring by 2061.[64] Given these survey results, even a very conservative estimate would assign at least a 0.1% chance to an AI-driven catastrophe as bad or worse than human extinction in the next 100 years.[65] Furthermore, it is plausible that $1 billion spent on AI safety would decrease the probability of such an outcome by at least 1%.[66] Consequently, $1 billion would provide at least a 0.001% absolute reduction in existential risk.[67] Thus, by donating $10,000 to AI safety, Shivani can provide a 1 in 10 million relative reduction and a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years.[68]

Shivani's options are as follows:

Shivani: Shivani has $10,000 to donate and she has four options:

i *Against Malaria Foundation* She saves in expectation 2.5 lives.

ii *Asteroid detection* She can provide a 1 in 33,000 trillion absolute reduction in the probability of extinction from an asteroid collision in the next 100 years.

iii *Pandemic prevention* She can provide a 1 in 50 trillion absolute reduction in the probability of an extinction-level pandemic in the next 100 years.

[64] Grace et al. (2018, p. 731).

[65] Greaves and MacAskill (2021, pp. 14–15).

[66] Greaves and MacAskill (2021, p. 15).

[67] Greaves and MacAskill (2021, p. 15).

[68] $0.01 \cdot 10000/10^9 = 10^{-7}$ (1 in 10 million). $0.001 \cdot 0.01 \cdot 10000/10^9 = 10^{-10}$ (1 in 10 billion).

iv *AI safety* She can provide a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years.

As mentioned earlier, we have assumed that Longtermism is true in Shivani's situation if and only if, in expectation, more than 2.5 additional lives exist in the far future if she donates to one of the longtermist causes. For it to be the case that over 2.5 additional lives exist in the far future if she donates to asteroid detection, the expected number of beings in the far future must be over 83,000 trillion.[69] Similarly, for it to be the case that over 2.5 additional lives exist in the far future if she donates to pandemic prevention, the expected number of beings in the far future must be over 125 trillion.[70] Finally, for it to be the case that over 2.5 additional lives exist in the far future if she donates to AI safety, the expected number of beings in the far future must be over 25 billion.[71] Is the expected number of lives in the far future large enough for Longtermism to be true in Shivani's situation?

Table 2

Expected Population Sizes Required for Longtermism

Asteroid detection	83,000 trillion
Pandemic prevention	125 trillion
AI safety	25 billion

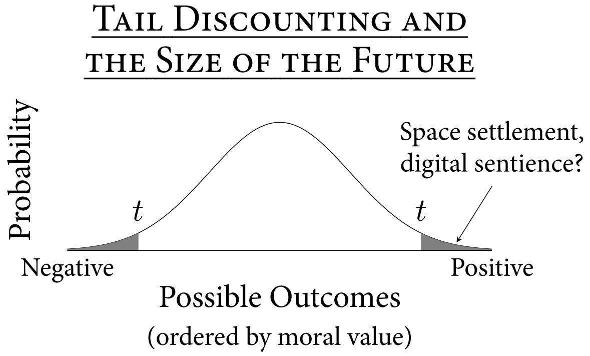
[69] $8.3 \cdot 10^{16} \cdot 3 \cdot 10^{-17} \approx 2.5.$

[70] $1.25 \cdot 10^{14} \cdot 2 \cdot 10^{-14} = 2.5.$

[71] $2.5 \cdot 10^{10} \cdot 10^{-10} = 2.5.$

3.2 Is the size of the future large enough?

Longtermism might depend on the possibilities of space settlement or the creation of digital minds because these possibilities inflate the value of the future; given these possibilities, the stakes are so high that even small reductions in existential risks have enormous expected value. If Longtermism depends on these possibilities, Tail Discounting undermines Longtermism if obtaining an outcome at least as good as these is very unlikely. In that case, Tail Discounting would ignore these possibilities, and the size of the future would not be large enough for Longtermism to be true (see the graph below).



Space settlement and the creation of digital minds might be the kind of unlikely best-case scenarios Tail Discounting ignores. However, it seems that the number of expected lives in the far future is sufficiently large for the argument for Longtermism to go through, even if we ignore these very-small-probability scenarios.[72] This is because there might be in expectation a sufficient number of individuals in

[72] Greaves and MacAskill (2021, §3).

the future if humanity survives for a long time on Earth. Based on the estimate of extinction risk due to natural causes, the expected future lifespan of humanity is at least 87,000 years.[73] On the other hand, the average lifespan of hominins is around one million years. Assuming a constant population size of 11 billion and an average lifespan of 80 years, this would mean that the expected number of humans is 12 trillion if humanity lives for a further 87,000 years and 140 trillion if humanity lives for a further million years.[74]

So, if humanity lives for 87,000 years in expectation, then AI safety leads to Longtermism (given that 12 trillion is greater than the required 25 billion expected future lives). This means that if Shivani donates to AI safety, more than 2.5 additional individuals live in the far future in expectation—so Longtermism is true in her situation. However, asteroid detection and pandemic prevention do not lead to Longtermism, as the expected number of individuals is not large enough (conditional on ignoring the very-small-probability scenarios). However, if humanity lives for one million years in expectation, then pandemic prevention also leads to Longtermism (given that 140 trillion is greater than the required 125 trillion expected future lives). In that case, more than 2.5 additional individuals live in the far future in expectation if Shivani donates to pandemic prevention.

However, humans are an atypical species, so extinction risk due to natural

[73] Snyder-Beattie et al. (2019).

[74] $11 \cdot 10^9 \cdot 87000/80 \approx 1.2 \cdot 10^{13}$ and $11 \cdot 10^9 \cdot 1000000/80 \approx 1.4 \cdot 10^{14}$. The UN Department of Economic and Social Affairs projects the world population to plateau at 11 billion. See United Nations and Social Affairs (2019). However, there are also signs of population decline. See Bricker and Ibbitson (2019). Note that the higher the world population is, the easier it is for Longtermism to be true; one antimalarial bednet will always save just one (or at most a few) people, but asteroid detection, pandemic prevention and AI safety will affect everyone.

causes and the lifespan of a typical hominin species may not be suitable bases for estimates of humanity's lifespan. How long might humanity survive? Even if we only stay on Earth, we have around one billion years until the Earth becomes uninhabitable.[75] If humanity survives for a billion years with a constant population size of 11 billion and an average lifespan of 80 years, then the number of humans would be 140,000 trillion.[76] In that case, asteroid detection, pandemic prevention and AI safety all would lead to Longtermism. But of course, humanity may become extinct well before the Earth becomes uninhabitable. How long must humanity's future be for asteroid detection, pandemic prevention and AI safety to lead to Longtermism?

For asteroid detection to lead to Longtermism, humanity's expected lifespan (ignoring the tail outcomes) must be at least 600 million years (given a constant population size of 11 billion and a human lifespan of 80 years).[77] Then, the expected number of humans in the far future is above the required 83,000 trillion. Pandemic prevention, in turn, leads to Longtermism if humanity's expected lifespan is at least 900,000 years (again, given a constant population size of 11 billion and a human lifespan of 80 years). Then, the expected number of future beings is above the required 125 trillion.[78]

Lastly, how long must humanity's future be for AI safety to lead to Longtermism? Suppose that the far future starts after 100 years. The expected number of

[75] Adams (2008). In principle, it might be possible to stay on Earth and keep the planet habitable for longer by changing its orbit or through stellar engineering projects that increase the sun's lifespan.

[76] $11 \cdot 10^9 \cdot 10^9/80 \approx 1.4 \cdot 10^{17}$.

[77] $11 \cdot 10^9 \cdot 604 \cdot 10^6/80 > 8.3 \cdot 10^{16}$.

[78] $11 \cdot 10^9 \cdot 909091/80 > 1.25 \cdot 10^{14}$.

beings in the far future is sufficiently large (above 25 billion) if humanity's expected lifespan *in the far future* is at least 182 years (given a constant population size of 11 billion and a human lifespan of 80 years).[79] Assuming a constant risk of extinction per year, this will be the case if humanity's expected lifespan is 265 years (this includes humanity's expected lifespan in the near and the far future). So, for AI safety to lead to Longtermism, it would have to be the case that humanity's expected lifespan is at least 265 years.

It seems plausible that humanity's expected lifespan is at least 265 years. This would be true if the risk of extinction per year is at most 0.38%.[80] Assuming a constant risk throughout the next 100 years, Ord's (2020, p. 167) estimate for existential risk is below this.[81] So, even if the probability of human extinction were 1/6 in the next 100 years, this would still be low enough for AI safety to lead to

[79] $11 \cdot 10^9 \cdot 182/80 > 2.5 \cdot 10^{10}$.

[80] $1/0.00377 \approx 265$. With a 0.00377 risk of extinction per year, humanity's expected number of years in the far future (after the next 100 years) is

$$1/0.00377 - \sum_{n=1}^{100}(1-0.00377)^n \approx 182.$$

This includes the possibility that humanity survives for a very long time, even when unlikely. However, these outcomes do not contribute much to the expectation. For example, the probability that humanity survives at least 2000 years is $(1-0.00377)^{2000} \approx 0.0005$—a probability that is plausibly above the discounting threshold. The contribution of the next 2000 years to humanity's expected lifespan is

$$\sum_{n=1}^{2000}(1-0.00377)^n \approx 264.$$

This is close to the expected lifespan of humanity (265 years).

[81] Existential risk in the next 100 years is 1/6 if the risk per year is 0.18% ($(1-0.0018)^{100} \approx 0.835$.) This is lower than the maximum 0.38% probability of human extinction per year with which AI safety leads to Longtermism. Ord (2020) does not give an estimate for extinction risk in the next 100 years. However, he believes this to be significantly lower than 1/6 (personal correspondence).

Longtermism. However, the probability of human extinction is lower than 1/6, as human extinction is just one type of existential catastrophe. Thus, the case for Longtermism from AI safety is even stronger.

Furthermore, there are many factors that we have not taken into account. First, it seems plausible that the risk of extinction per year is not constant.[82] For example, there may be a few particularly dangerous moments expected to happen within the next couple of centuries, such as the development of artificial general intelligence, after which the yearly risk of extinction is significantly lower.[83] If we now live in a 'time of perils' after which the yearly risk of extinction is significantly lower, existen-

[82] According to the "Simple Model" of existential risk mitigation, the expected value of the future is

$$EU[F] = v \sum_{n=1}^{\infty} (1-r)^i = \frac{v(1-r)}{r},$$

where v is the value of human existence each century (assumed to be constant) and r is a per-century existential risk (also assumed constant). In this model, the value of the future is the value of a single century divided by the per-century risk. See Ord (2020, appendix E). This model implies that the value of reducing existential risk this century by some fraction f is $EU(X) = fv$. This result is surprising because the value of existential risk reduction is capped at the value v (an additional century of human existence)—it is not astronomical. See Thorstad (n.d.) for a discussion of the Simple Model. If human population stays at a constant 11 billion, each person living for 80 years, then the value of an additional century of human existence (measured in lives) is approximately 14 billion ($1.25 \cdot 11,000,000,000 = 13,750,000,000$). It was assumed that Shivani could save in expectation 2.5 lives by donating to Against Malaria Foundation. So, if the Simple Model is right, Shivani should donate to a longtermist cause if she can decrease (relatively) the probability of extinction by at least 1 in 5 billion ($2.5/13,750,000,000 \approx 2 \cdot 10^{-10}$).

[83] Thorstad (n.d.) argues that the belief that existential risks are high is unlikely to ground the overwhelming importance of existential risk mitigation unless coupled with the time of perils hypothesis. This is so because the higher the probability of existential risk per century, the shorter the expected lifespan of humanity is. Therefore, a high level of risk means the size of the future is correspondingly smaller. However, if we now live in a particularly dangerous period after which existential risk is much lower, then the size of the future can be considerable. However, Thorstad (n.d.) also argues that the time of perils hypothesis is probably false. Therefore, pessimism about existential risks does not justify the overwhelming importance of existential risk mitigation.

tial risk mitigation more easily leads to Longtermism.[84] Ord (2020, pp. 189–191) argues that humanity's first task is to reach existential security—a place where existential risk is low and stays low.

The size of the future seems large enough for Longtermism to be true—even if we ignore very-small-probability scenarios such as space settlement and digital minds. Asteroid detection leads to Longtermism if humanity's expected lifespan is at least 600 million years. With pandemic prevention and AI safety, the required expected lifespans are 900,000 and 265 years, respectively. Finally, it would be overconfident to be near-certain that space settlement or digital sentience will not occur, given that there is no known reason why they should be physically impossible.[85] If one gives a non-negligible probability for at least one of these scenarios, then the expected number of lives in the far future will be much greater. To conclude, even if we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of lives in the far future seems large enough for Longtermism to be true. Thus, the Small Future Argument does not undermine Longtermism.

[84] The astronomer Sagan (1997, p. 173) writes about the time of perils: "Some planetary civilizations see their way through, place limits on what may and what must not be done, and safely pass through the time of perils. Others are not so lucky or so prudent, perish." Rees (2003, pp. 7-8) echoes this by writing that "the most crucial location in space and time (apart from the big bang itself) could be here and now. [...] What happens here on Earth, in this century, could conceivably make the difference between a near eternity filled with ever more complex and subtle forms of life and one filled with nothing but base matter."

[85] The entrepreneur Elon Musk (n.d.) wants humanity to be a spacefaring civilization: "You want to wake up in the morning and think the future is going to be great—and that's what being a spacefaring civilization is all about. It's about believing in the future and thinking that the future will be better than the past. And I can't think of anything more exciting than going out there and being among the stars."

4 Probability of making a difference

This section discusses the argument that the probability of making a difference to whether or not an existential catastrophe occurs is tiny, and thus, we should ignore the possibility of influencing the occurrence of existential catastrophes. One type of Probability Discounting naturally captures this idea. However, I will show that the different versions of this view violate Statewise Dominance or Acyclicity, which makes them less plausible as theories of instrumental rationality.

4.1 State Discounting

The final objection to Longtermism from discounting small probabilities is that the probability of making a difference to whether or not an existential catastrophe occurs is so tiny that it should be discounted down to zero—let's call this the *No Difference Argument*.

> **No Difference Argument:** The probability of making a difference to whether or not an existential catastrophe occurs is so small that we should ignore the possibility of making a difference.

If it is indeed the case that Shivani has only a negligible probability of having an impact with all of the possible longtermist causes, and such small probabilities should be discounted down to zero, then she should instead donate to the Against Malaria Foundation. Consequently, Longtermism would be false in her situation.

Recall that the absolute reductions in the probability of extinction that Shivani can provide are 1 in 33,000 trillion with asteroid detection, 1 in 50 trillion with pandemic prevention and 1 in 10 billion with AI safety (see table 3). If Shivani plans to donate less than $10,000, her probability of impact is even smaller.[86] As these numbers are tiny, it may not be unreasonable to ignore the possibility of Shivani making a difference to existential risks with her donation to the longtermist causes.[87] But which version of Probability Discounting allows her to do this?

Table 3

Absolute Reductions of Extinction Risks with $10,000

Asteroid detection	1 in 33,000 trillion
Pandemic prevention	1 in 50 trillion
AI safety	1 in 10 billion

One version of Probability Discounting captures the No Difference Argument naturally. Recall that Naive and Tail Discounting ignore outcomes associated with small probabilities. However, one might ignore *states* associated with small probabilities instead—let's call this *State Discounting*.[88]

State Discounting For all prospects X and Y, $X \succsim Y$ if and only if

[86] Conversely, if she plans to donate more than $10,000, her probability of impact is higher. It is plausible that at least some individuals are in a position to have a non-negligible impact on existential and extinction risks via donations. Sam Bankman-Fried, the founder and CEO of FTX and a member of Giving What We Can, set out to make as much money as he could in order to give away everything he earned to charity. He is now the primary funder of the FTX Foundation's Future Fund, which works to improve humanity's odds of surviving and flourishing for thousands of years or longer. See FTX Future Fund (2022).

[87] Note that some might have a non-negligible impact on existential risks by doing direct work instead of donating money. For them, Longtermism may be true in the context of choosing which career to pursue or how to spend one's free time.

[88] Note that the definition of State Discounting given here considers very-small-probability states in cases where the prospects would otherwise have equal probability-discounted expected utility.

- $EU(X)_{pd} > EU(Y)_{pd}$ or
- $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,

where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that no state of negligible probability occurs.

In order to use State Discounting to argue against Longtermism, we need a way of individuating states that guarantees that states in which Shivani makes a difference to existential risks are negligible. This can be done by individuating states in terms of whether some act makes a difference to existential catastrophes as follows: In one state, an existential catastrophe happens no matter what one does; in another state, one's actions make a difference to whether or not the catastrophe happens; and in the final state, an existential catastrophe does not happen no matter what one does. Let's call the second state a *difference-making state*. If the difference-making state is associated with a tiny probability, then one should ignore it. In effect, one would then ignore the possibility of making a difference to whether or not an existential catastrophe happens.

There are different ways of partitioning states, and thus, many versions of State Discounting. The focus of this section will be a version of State Discounting on which states are partitioned by comparing prospects to some status quo prospect, which corresponds to doing nothing.[89] Let's call this view *Baseline State Discounting*.

[89] On another version of State Discounting, prospects are always compared two at a time, and the possible states of the world are partitioned for every pairwise comparison separately. On a third version, states are partitioned by comparing all available options at once. See the appendix for a discussion of these two views.

> **Baseline State Discounting:** States are partitioned by comparing every prospect to a status quo prospect (each separately).

How might Baseline State Discounting undermine Longtermism? Recall that by donating $10,000 to the Against Malaria Foundation, Shivani can save 2.5 lives in expectation—let's round that to 2. By donating the same money to AI safety, she can provide a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years. Baseline State Discounting compares the Against Malaria Foundation and AI safety to a status quo prospect (i.e., 'do nothing'), each separately.

Let's start by comparing AI safety to doing nothing. In order to capture the idea of the No Difference Argument, states must be individuated based on whether Shivani makes a difference to an AI-driven catastrophe as follows (see table 4): In state 1, an AI causes an existential catastrophe no matter what Shivani does. In state 2, an AI does not cause an existential catastrophe if she donates to AI safety, but it will cause an existential catastrophe if she does nothing. Lastly, in state 3, an AI does not cause an existential catastrophe no matter what she does. If Shivani's discounting threshold is higher than 1 in 10 billion, then she should ignore the possibility of state 2 obtaining. Consequently, the probability-discounted expected utility of AI safety equals (or is marginally better than) that of doing nothing. In effect, Shivani would then ignore the possibility of making a difference to whether or not an AI-driven existential catastrophe happens.

Table 4

AI Safety vs. Baseline

	State 1 $p \approx 0.001$	**State 2** $p = 10^{-10}$	**State 3** $p \approx 0.999$
AI safety	AI doom	No AI doom	No AI doom
Do nothing	AI doom	AI doom	No AI doom

Donating to the Against Malaria Foundation involves no uncertainty, as (we have assumed) it certainly saves two lives. As the Against Malaria Foundation certainly results in a better outcome than doing nothing, its probability-discounted expected utility is greater than that of doing nothing (see table 5).

Table 5

AMF vs. Baseline

	State 1
AMF	Two additional lives saved
Do nothing	No additional lives saved

So, the probability-discounted expected utility of AI safety equals that of doing nothing, while the probability-discounted expected utility of the Against Malaria Foundation is greater than that. Therefore, Shivani should donate to the Against Malaria Foundation, and Longtermism is false in her situation. Thus, Baseline State Discounting provides a prima facie case against Longtermism. If states are partitioned as in table 4, and the difference-making state (i.e., state 2) has negligible probability with all of the possible longtermist causes, then Baseline State Discounting undermines Longtermism.

4.2 State Individuation Problem

However, State Discounting faces an analogous problem to the Outcome Individuation Problem but with states instead of outcomes:[90]

> **State Individuation Problem:** If one individuates states with too much detail, all states have negligible probabilities. Is there a privileged way of individuating states that avoids this?

Earlier, states were individuated in terms of whether or not Shivani could make a difference to the occurrence of an AI-driven catastrophe. However, there are many ways in which such a catastrophe might happen. The occurrence of an AI-driven catastrophe was treated as a privileged basis for individuating states. We were interested in whether Shivani can affect the occurrence of an AI-driven catastrophe with no regard to how it might happen or how much utility is at stake. However, this seems arbitrary. Why should states be individuated in this way rather than some other way?

Apart from individuating states as finely as possible, it seems the only non-arbitrary way of individuating states is by the utilities of their outcomes. But in the case of existential risk from an AI, individuating states by the utilities of their outcomes would most likely result in many different states instead of just three,

[90] See §3 in Chapter 4 of this thesis.

as in the earlier example (table 4). This is so because these catastrophic scenarios would most likely differ in value. As a result, individuating states by the utilities of their outcomes does not guarantee that one will ignore the possibility of influencing the occurrence of an AI-driven catastrophe if and only if its probability is tiny. For example, one might ignore the possibility of making a difference even when the probability of doing so is high. This can happen if the different scenarios in which one makes a difference differ in value, and all these scenarios have tiny probabilities (even though their total probability is high). So, individuating states by the utilities of their outcomes does not capture the idea of the No Difference Argument.

Furthermore, individuating states by the utilities of their outcomes results in a violation of dominance.[91] Let $X \succ Y$ mean that X is strictly preferred to Y. Then, Baseline State Discounting violates the following dominance principle if states are individuated by utilities:[92]

> **Statewise Dominance:** If the outcome of prospect X is at least as preferred as the outcome of prospect Y in all states, then $X \succsim Y$. Furthermore, if in addition the outcome of X is strictly preferred to the outcome of Y in some possible state, then $X \succ Y$.

To see how Baseline State Discounting violates Statewise Dominance if states are individuated by utilities, consider the following prospects:

> **Space Settlement:** The Earth has a billion years left until the Sun expands and makes the Earth uninhabitable. However, a space settle-

[91] See §3 in Chapter 4 of this thesis.

[92] Savage (1951, p. 58) and Luce and Raiffa (1957, p. 287).

ment program might expand humanity's lifespan, with some cost ϵ. There are two alternative programs whose successes depend on some mutually exclusive events E_1, E_2, E_3 and E_4 as follows:[93]

> *Space program 1* Gives a 3% chance of humanity surviving for two billion years (if event E_1 happens) and a 2% chance of humanity surviving for five billion years (if event E_2 or E_3 happens). Otherwise, humanity will survive for a billion years on Earth (if event E_4 happens).
>
> *Space program 2* Gives a 4% chance of humanity surviving for two billion years (if event E_1 or E_2 happens) and a 1% chance of humanity surviving for five billion years (if event E_3 happens). Otherwise, humanity will survive for a billion years on Earth (if event E_4 happens).

Suppose the discounting threshold is (implausibly) just above 2%, and also suppose that the utility of humanity's lifespan equals its duration (in billions of years). Let's first compare *Space program 1* to the baseline, which again is 'do nothing.' Individuating states by the utilities of their outcomes results in the following states (see table 6): In state 1, humanity lives for two billion years if *Space program 1* is chosen, and otherwise, humanity lives for one billion years; in state 2, humanity lives for five billion years if *Space program 1* is chosen, and otherwise, humanity

[93] Note that usually in decision theory an event is defined as a set of states, which is not the case here. For example, state 2 in table 6 is composed of two mutually exclusive events. Here 'event' is used in its common meaning outside of decision theory. 'State', in turn, refers to a collection of maximally fine-grained possible states of the world. The reason for understanding states in this less fine-grained way is that maximally fine-grained states would all have probabilities below the discounting threshold.

lives for one billion years; and in state 3, both *Space program 1* and doing nothing result in humanity living for one billion years. The probability of state 2 is below the discounting threshold, so one should conditionalize on state 2 not happening. Then, the probability-discounted expected utility of *Space program 1* is 1.03—ϵ.[94]

Table 6

Space Program 1 vs. Doing Nothing

	State 1	**State 2**	**State 3**
Event	E_1	E_2 or E_3	E_4
p	0.03	0.02	0.95
Space program 1	2—ϵ	5—ϵ	1—ϵ
Do nothing	1	1	1

Next, let's compare *Space program 2* to the baseline. In this case, states are individuated similarly as before, except that the states have slightly different probabilities, as now E_2 results in state 1* (see table 7). As before, state 2* has negligible probability, so the possibility of state 2* is ignored. Consequently, the probability-discounted expected utility of *Space program 2* is 1.04—ϵ.[95]

[94] $0.03/0.98 \cdot (2 - \epsilon) + 0.95/0.98 \cdot (1 - \epsilon) \approx 1.03 - \epsilon.$

[95] $0.04/0.99 \cdot (2 - \epsilon) + 0.95/0.99 \cdot (1 - \epsilon) \approx 1.04 - \epsilon.$

Table 7

Space Program 2 vs. Doing Nothing

	State 1*	**State 2***	**State 3***
Event	E_1 or E_2	E_3	E_4
p	0.04	0.01	0.95
Space program 2	$2{-}\epsilon$	$5{-}\epsilon$	$1{-}\epsilon$
Do nothing	1	1	1

The probability-discounted expected utility of *Space program 2* is greater than that of *Space program 1* (1.04—ϵ vs. 1.03—ϵ), so *Space program 2* is better than *Space program 1*. However, the only difference between these alternatives is that the former results in a lifespan of two billion years for humanity if event E_2 happens, while the latter results in a lifespan of five billion years in that case. So, when states are individuated in the usual way (see table 8), the two space programs give the same outcomes in states 1**, 3** and 4**, but *Space program 1* gives a better outcome in state 2**. This is a violation of Statewise Dominance. This Statewise Dominance violation happens because the partition of states is different for each option, leading to a situation where the states in which Space Program 1 beats Space Program 2 are ignored for Space Program 1 but not for Space Program 2. So, the most plausible way of individuating states (i.e., by utilities) leads to a violation of Statewise Dominance—which makes Baseline State Discounting less plausible as a theory of instrumental rationality.

Table 8

Space Program 1 vs. Space Program 2

	State 1**	**State 2****	**State 3****	**State 4****
Event	E_1	E_2	E_3	E_4
p	0.03	0.01	0.01	0.95
Space program 1	$2-\epsilon$	$5-\epsilon$	$5-\epsilon$	$1-\epsilon$
Space program 2	$2-\epsilon$	$2-\epsilon$	$5-\epsilon$	$1-\epsilon$

To summarize, the No Difference Argument states that the probability of making a difference to whether or not an existential catastrophe happens is so tiny that the possibility of making a difference should be ignored. Baseline State Discounting captures this idea naturally. And, it presents a prima facie challenge to Longtermism, as there is only a tiny probability that Shivani can make a difference to whether or not an existential catastrophe occurs. However, Baseline State Discounting faces the State Individuation Problem. As before, one might solve this by individuating states by the utilities of their outcomes. But if states are individuated by utilities, then it is not guaranteed that Baseline State Discounting ignores the possibility of making a difference if and only if the probability of doing so is tiny. So, Baseline State Discounting does not capture the idea behind the No Difference Argument if states are individuated by utilities. Furthermore, individuating states by the utilities of their outcomes also results in a violation of Statewise Dominance, which makes Baseline State Discounting less plausible as a theory of instrumental rationality. Nevertheless, one might still insist that there is some *other* privileged way of individuating states that avoids the violation of Statewise Dominance. Alternatively, one might reject Baseline State Discounting and cash out the No Difference Argument in some other way. So, the No Difference Argument might still challenge Longtermism. However, the next section presents a more general response to the No Difference Argument.

5 Probability Discounting and Each-We Dilemmas

This section argues that Probability Discounting faces Each-We Dilemmas. These can be solved by accepting *Collective Difference-Making*. However, doing so also blocks the No Difference Argument. Some possible justifications for Collective Difference-Making will be discussed.

5.1 Collective Difference-Making

According to Parfit (1984, p. 91), a theory faces Each-We Dilemmas if "there might be cases where, if each does better in this theory's terms, we do worse, and vice versa."[96] To see how Probability Discounting faces Each-We Dilemmas, consider the following case (see table 9 depicting the decision-situation faced by a single agent):[97]

[96] Each-We Dilemmas differ from Prisoner's Dilemmas because in the former even impartial and altruistic agents who accept the same moral theory can end up choosing worse options by the lights of that theory when those choices are evaluated together.

[97] Versions of Probability Discounting that ignore very-small-probability outcomes face the following Each-We Dilemma:

> **Asteroid:** Multiple asteroids are heading toward the Earth, and for each of them, there is a tiny probability that it will hit unless it is stopped. However, the probability that at least one of them will hit the Earth is high if none of the asteroids are stopped.

> **Asteroid:** An asteroid is heading toward the Earth and will almost certainly hit unless stopped. There are multiple asteroid defense systems, and (unrealistically) each has a tiny probability of hitting the asteroid and preventing a catastrophe. However, the probability that one of them succeeds is high if enough of them try. Attempting to stop the asteroid involves some small cost ϵ.[98]

Table 9

Asteroid

	State 1	**State 2**	**State 3**
Attempt	Collision $-\epsilon$	No collision $-\epsilon$	No collision $-\epsilon$
Do nothing	Collision	Collision	No collision

In this case, the probability of state 2 happening is below the discounting threshold, so the possibility of state 2 should be ignored. However, then doing nothing is better than attempting to stop the asteroid because it gives a better outcome in states 1 and 3. So, Probability Discounting recommends against attempting to stop the asteroid because the probability of making a difference is below the discounting threshold, and trying to stop the asteroid incurs a small cost. Consequently,

> There are multiple asteroid defense systems, and each can only target one asteroid. Attempting to stop an asteroid involves some small cost ϵ.

As the agents can only attempt to stop one asteroid, and the probability of this asteroid hitting the Earth is tiny, versions of Probability Discounting that ignore very-small-probability outcomes recommend against attempting to stop the asteroid. Consequently, an asteroid will almost certainly hit the Earth—which could have been prevented had enough agents attempted to do so.

[98] This cost is so small that the asteroid hitting the Earth is worse than a cost of ϵ to all the relevant people.

the asteroid will almost certainly hit the Earth—which could have been prevented almost certainly had enough agents attempted to do so.

Many have appealed to expected benefits in order to solve collective action problems.[99] For example, it is sometimes argued that one cannot justify voting by merely appealing to the consequences of one's act because there is only a minuscule probability that one vote makes a difference.[100] The expected benefits of voting can nonetheless be great because if one's vote makes a difference, it will impact millions of people.[101] However, if one ought to discount very small probabilities, then appealing to expected benefits cannot solve collective action problems in which it is almost certain of each person that they make no difference. If one vote is extremely unlikely to make a difference, and one should ignore tiny probabilities, then the expected benefits of voting are negligible.

If Probability Discounting is to avoid Each-We Dilemmas, agents must somehow take into account the choices of other people. They must accept

> **Collective Difference-Making:** One ought to take into account the choices of other people and consider whether the collective has a non-negligible probability of making a difference.[102]

[99] See Parfit (1984, pp. 73–75), Parfit (1988) and Kagan (2011). For a criticism of this solution, see Nefsky (2011).

[100] Parfit (1984, p. 73).

[101] Parfit (1984, pp. 73–75).

[102] Note that, on Collective Difference-Making, it matters whether the small probabilities are independent for the different agents. Suppose that a googolplex agents face *Pascal's Mugging*. The probability that at least one of them gets a thousand quadrillion happy days in the Seventh Dimension is still small even if they all pay the mugger because the probability of obtaining the great outcome is not independent for the different agents: Either the mugger has magical powers, or he does not. So, Collective Difference-Making recommends that the agents ignore the small prob-

There are several different ways to interpret Collective Difference-Making. On one interpretation, agents should choose a small enough discounting threshold so that Each-We Dilemmas do not arise to begin with (and adjust the threshold lower if they anyway do arise). This interpretation is 'collective' because agents ought to take into account the choices of others when choosing the discounting threshold. On another interpretation, all the choices faced by different agents should be evaluated collectively, and if the total probability of some event or outcome is above the discounting threshold, then no one should discount. This latter view is similar to what Monton (2019) and Smith (2016) say in diachronic cases, where we consider different choices made by the same agent over time. They argue that relevantly similar choices faced by one individual must be evaluated collectively, and one should not discount if the total probability of some event or outcome is above the discounting threshold.[103] So, on this interpretation, Collective Difference-Making implies that one should reason as if one was facing sequentially all the choices faced by different agents.

The probability that Shivani and all the other agents together can make a difference to existential risks seems non-negligible. For example, if we spend $1 billion on AI safety, it is plausible that we can provide at least a 1 in 100,000 absolute reduction in the probability of an AI-driven catastrophe.[104] This estimate is

ability. However, if the probabilities were independent, then Collective Difference-Making would recommend against discounting, provided that the total probability of at least one person obtaining the great outcome is sufficiently high.

[103] The approach advocated by Monton (2019) and Smith (2016) assumes that there are no intrapersonal Each-We Dilemmas because rational agents have the power to commit to making some choices in the future.

[104] $0.001 \cdot 0.01 = 0.00001$. Greaves and MacAskill (2021, pp. 14–15) estimate that there is at

conservative. As mentioned earlier, the median expert estimate for an AI-driven catastrophe at any point in time is 5%, while the calculation assumed a 0.1% risk in the next 100 years. Also, $1 billion spent on AI safety might decrease the probability of an AI-driven catastrophe by more than 1%. So, if one ought to accept Collective Difference-Making, then—plausibly—Probability Discounting does not undermine Longtermism. Shivani should not ignore the possibility of making a difference because she and the other agents have a non-negligible chance of preventing an existential catastrophe.

The details of Collective Difference-Making do not matter for the purposes of this chapter, so I will only briefly mention some possible justifications for and problems with Collective Difference-Making. The details do not matter because, if Collective Difference-Making is plausible, then Probability Discounting does not undermine Longtermism, as Shivani and all the other agents have a non-negligible chance of making a difference. But if Collective Difference-Making is implausible, then Probability Discounting faces Each-We Dilemmas, making it implausible as well. Either way, Probability Discounting does not undermine Longtermism.

5.2 Justifications for Collective Difference-Making

How can Collective Difference-Making be justified? In response to collective action problems, some argue that we have reasons for action coming from the participatory nature of one's act. On these views, the reason for action is that by doing

least a 0.1% chance of an AI-driven catastrophe in the next 100 years, and that $1 billion of spending would decrease this probability by at least 1%. See Greaves and MacAskill (2021, p. 15).

so, one could be part of a group of people who together could make a difference.[105] For example, some argue that we have collective reasons for action.[106] On this view, groups, like individuals, have reasons to make outcomes better, benefit other people, avoid harming other people and benefit themselves. We have reasons as a group to carry out some action because we would together be making things better.[107] Furthermore, there might be things that some groups ought to do, even if they have never coordinated in the past nor will ever coordinate in the future.[108] This view can solve collective action problems if the reasons of groups bear on the reasons of individuals. In that case, the agents in *Asteroid* may have a collective reason to attempt to stop the asteroid and an individual reason to do their part. Similarly, Shivani and the other agents may have a collective reason to prevent an existential catastrophe (if they have a non-negligible probability of having an impact) and an individual reason to do their part.

Others, in turn, argue that one's act can be part of causing some outcome without making a difference.[109] This can happen when the outcome not happening would be at least partly a result of there not having been enough similar acts.[110]

[105] Nefsky (2017, p. 2756). For a criticism of these views, see Nefsky (2015).

[106] See for example Dietz (2016). Consider also this view from Parfit (1984, p. 70):

> "Even if an act harms no one, this act may be wrong because it is one of a set of acts that together harm other people. Similarly, even if some act benefits no one, it can be what someone ought to do, because it is one of a set of acts that together benefit other people."

See also Parfit (1984, pp. 31–31).

[107] Dietz (2016, p. 960).

[108] Dietz (2016, p. 957).

[109] Nefsky (2017).

[110] Nefsky (2017, p. 2753).

The idea is that one has a reason to act in a certain way because one could be making a causal contribution toward bringing about some outcome (even though one would not make a difference in expectation). The conditions for making a causal contribution without making a difference are that it is up in the air whether or not the outcome in question will occur; that part of what could determine whether it occurs is whether enough people act in the relevant way going forward; and that it is up in the air whether or not enough people will act in that way going forward.[111]

On this view, the agents in *Asteroid* should attempt to stop the asteroid because doing so might be making a causal contribution toward stopping it, even though in expectation they would not be making a difference.[112] Similarly, Shivani should mitigate existential risks because she might thereby be making a causal contribution toward preventing an existential catastrophe (even though in expectation she would not be making a difference). In both Shivani's case and *Asteroid*, it is up in the air whether or not the existential catastrophe will occur; part of what could determine whether it occurs is whether enough people mitigate existential risks; and

[111] Nefsky (2017, p. 2758). On reasons to vote, Nefsky (2017, pp. 2756–2757) writes: "But, contrary to the expected utility approach, the main reason to vote does not come from this minuscule chance of making a difference—from extremely remote chance of the election turning on your vote. Rather, it comes from the fact that your vote could help to elect Mr. Powers [the better candidate] regardless of whether the election turns on it (which it almost certainly will not). Your vote could help because, at the time at which you vote, more votes for Mr. Powers are needed to prevent the disastrous outcome, and there is no guarantee that there will be enough such votes. So, by voting, you are making a causal contribution toward preventing the bad outcome, when there is a real risk that this outcome will not be prevented due to a lack of exactly that sort of contribution. Making such a contribution in those circumstances makes progress toward preventing the bad outcome, even if what happens will not turn on your having done so."

[112] It is unclear whether Nefsky would apply this theory to cases such as *Asteroid*. On cases in which each person has a tiny chance of triggering some result regardless of what others do (such as *Asteroid*), Nefsky (2011, p. 367n11) writes: "It seems to me, though, that such a case would not be a collective harm case."

it is up in the air whether or not enough people will mitigate existential risks.

Alternatively, one can also justify Collective Difference-Making with, for example, rule-consequentialism. Rule-consequentialism states that agents should decide what to do by applying rules whose acceptance will produce the best consequences. Rule-consequentialism would (presumably) advise that the agents attempt to stop the asteroid because doing so conforms to a rule whose acceptance produces the best consequences. Similarly, Rule-consequentialism would (presumably) advise Shivani to mitigate existential risks because 'mitigate existential risks' is a rule whose acceptance produces the best consequences in the long run.

Another way of justifying something close to Collective Difference-Making comes from Evidential Decision Theory. According to Evidential Decision Theory, the best act is the one that gives the best expectations for the outcomes, conditional on one choosing it. Evidential Decision Theory is often contrasted with Causal Decision Theory. According to Causal Decision Theory, agents ought to maximize the best expected causal consequences. On this view, causality plays an important role in instrumental rationality: Only those consequences that have a causal link with one's act count. In contrast, evidentialists do not require a belief in a causal link between one's act and the consequences.[113]

Evidential Decision Theory favors something akin to Collective Difference-Making because it implies that an agent ought to reason as if they were choosing on behalf of all relevantly similar agents.[114] Evidential Decision Theory recom-

[113] See Nozick (1969).

[114] MacAskill et al. (2021) argue that an altruistic and morally motivated agent who is uncertain between Evidential and Causal Decision Theory should generally act following the former, even if

mends not discounting the probability of making a difference in *Asteroid* if doing so provides sufficient evidence of others also not discounting. And it may, if others are similar to the agent in relevant ways. The idea is that under certain conditions, conditional on some agent acting, it is likely that enough people act to deflect the asteroid, so the probability of stopping the asteroid is non-negligible. Similarly, Evidential Decision Theory recommends Shivani to mitigate existential risks if doing so provides sufficient evidence of others mitigating these risks as well. And it may, if others are similar to Shivani in relevant ways. However, Evidential Decision Theory does not solve Each-We Dilemmas in cases where one's actions do not provide suitably strong evidence of how other agents will act. If ignoring the small chance of stopping the asteroid does not provide sufficiently strong evidence of other agents doing so as well, then Evidential Decision Theory recommends doing nothing instead of attempting to stop the asteroid.

5.3 Problems with Collective Difference-Making

I have discussed some ways of justifying Collective Difference-Making. However, Collective Difference-Making faces some problems as well. First, to even start estimating the number of very-small-probability choices all agents make, one needs to know who counts as an agent. Do small children count? What about animals? Or possible intelligent aliens or AI? Evidential Decision Theory can solve this: All agents who are relevantly similar to oneself count (in proportion to how similar

she has a higher credence in the latter. They argue that the existence of correlated decision-makers will affect the stakes for Evidential Decision Theory but not for Causal Decision Theory and that it is rational to hedge if one faces decision-theoretic uncertainty.

they are to oneself) because then one's actions are evidence of how they will act. Another possible solution is that those on a collective endeavor with oneself count.[115] On this view, for example causally disconnected intelligent aliens do not count.

Another problem for Collective Difference-Making is the violation of Separability. Let X be a prospect that concerns what is going on in the part of the world we might make any difference to, and let Y be a prospect that concerns what happens somewhere far away, such as a distant galaxy. Also, let $X \oplus Y$ be the combined prospect of the near prospect X and the far prospect Y. Then, Separability states the following:[116]

Separability:

i For all near prospects X and Y, and any far prospect Z, $X \succ Y$ if and only if $X \oplus Z \succ Y \oplus Z$.

ii For all far prospects X and Y, and any near prospect Z, $X \succ Y$ if and only if $Z \oplus X \succ Z \oplus Y$.

Collective Difference-Making violates Separability because what one ought to do

[115] For example, Kutz (2000, p. 89) writes:"Jointly acting groups consist of individuals who intend to contribute to a collective end."

[116] Russell (2021, p. 15). Contrast Separability with *Background Independence*:

> **Background Independence:** For all prospects X and Y, and any far outcome z, $X \succ Y$ if and only if $X \oplus z \succ Y \oplus z$ (Russell, 2021, p. 18).

Background Independence is related to the Egyptology objection to the Average View in population ethics. See McMahan (1981, p. 115) and Parfit (1984, p. 420). The background outcome z does not add any uncertainty, so it will not interact with X and Y in different ways in different states. Thus, unlike Separability, Background Independence is consistent with (first-order) Stochastic Dominance. See Russell (2021, p. 18n13).

depends on what choices other distant agents face.[117] For example, Collective Difference-Making implies that the agents in *Asteroid* should not attempt to stop the asteroid if no other agents were facing the same choice; but given that enough others are also facing this choice, they should attempt to stop the asteroid. So, what agents should do depends on what choices others face.

Furthermore, there is a trade-off between maintaining Separability and avoiding Each-We Dilemmas. The fewer agents' choices one considers in one's decision-making, the more Each-We Dilemmas occur, and vice versa. For example, if one only takes into account the choices of other humans living on Earth right now, then one might end up in an Each-We Dilemma situation with future generations. Alternatively, if one only takes into account the choices of those who are on a collective endeavor with oneself, then one might end up in an Each-We Dilemma with those not on this collective endeavor.

Suppose that possible intelligent aliens would not be on a collective endeavor with us. We might then end up in the following kind of Each-We Dilemma with them:

[117] Wilkinson (2022, §6) shows that denying Probability Fanaticism leads to violations of Separability (or first-order Stochastic Dominance), even in cases where the choices of different individuals are probabilistically independent. See also Beckstead and Thomas (2020). However, Russell (2021) shows that (first-order) Stochastic Dominance and Separability are inconsistent (assuming *Positive Compensation*: One can always compensate for making things worse nearby by making things sufficiently better far away, and vice versa). Also see Goodsell (2021). Russell (2021, p. 14) writes: "[W]hat is better than what really does depend in strange ways on what is going on in distant space and time… it matters whether you think there is another St. Petersburg population lottery going on in a distant galaxy. This is bizarre—but Stochastic Dominance tells us that it is true." Stochastic Dominance is consistent with the separability of simple prospects, that is, prospects that have finitely many possible outcomes (Russell, 2021, p. 14). However, as Russell points out, whatever justifies the separability of simple prospects will probably also justify full Separability.

> **Asteroid 2:** Asteroids are heading toward different planets (one for each planet), and they will almost certainly hit unless they are stopped. There is one asteroid defense system on every planet, and (unrealistically) each has a tiny probability of hitting the asteroid and preventing a catastrophe. However, the probability that at least one of them hits an asteroid is high if enough of them try. Again, trying to stop the asteroid incurs some small cost ϵ.

It would be better if everyone attempted to stop the asteroid heading toward their planet. Probably at least one of the planets would survive. However, if one should ignore what happens on faraway planets, then one should ignore the possibility of successfully stopping the asteroid heading toward one's planet. Consequently, no planets survive. So, if one ignores the choices of some group of agents, then one might end up in an Each-We Dilemma with this group. On the other hand, if one cares about the difference all agents can make, then violations of Separability will be more common. Also, if there is a large number of agents, one might not discount tiny probabilities very often, if ever.[118]

Another problem for Collective Difference-Making is cluelessness: It seems impossible to evaluate how many very-small-probability choices other agents face. So, Collective Difference-Making needs some way of handling situations where

[118] Wilkinson (2022) writes on the long-run argument for maximizing expected value: "How well the world as a whole goes is not determined by just a few decisions by a single agent, but instead by countless different agents making separate small-scale decisions. In this setting, having all of those agents maximize expected value seems to be quite a good policy, even when doing so produces fanatical verdicts. Repeated enough times, even fanatical choices will pay off eventually." However, note that this will only happen if the probabilities are sufficiently independent for the different agents.

one is clueless about what choices others face. However, many other theories also face the problem of cluelessness, so this problem need not disadvantage Collective Difference-Making over the alternatives.[119]

Finally, another task for the proponents of Collective Difference-Making is to spell out the details of when agents should refrain from discounting small probabilities. Does it only have to be the case that sufficiently many agents face sufficiently many very-small-probability choices, or do enough of those agents also need to refrain from discounting? Do their choices need to be relevantly similar (such as attempts to stop a particular asteroid heading toward the Earth), or is it enough that they involve similarly small probabilities but in very different contexts? What happens if different agents assign different probabilities to the same events?

I will not attempt to solve these problems in this chapter. Instead, as mentioned earlier, my argument is that if Collective Difference-Making is implausible, then Probability Discounting is also implausible because it leads to Each-We Dilemmas. On the other hand, if Collective Difference-Making is plausible, then Probability Discounting does not undermine Longtermism because Shivani and all the other agents together have a non-negligible probability of making a difference. Either way, Probability Discounting and the No Difference Argument do not undermine Longtermism. However, discounting small probabilities might still be relevant to

[119] However, this problem may be more serious for Collective Difference-Making. For example, an agent might think there is a tiny probability that countless agents face very-small-probability choices. Should the agent discount that probability down to zero and ignore this possibility? If the agent ignores this possibility, then the number of individuals is small, and they are right to ignore it. On the other hand, if the agent does not ignore this possibility, then the number of individuals is large, and the agent is right not to ignore it.

what longtermists should focus on, as there might be a class of existential risks that we cannot make a difference to, even together.

6 Conclusion

I have discussed three arguments against Longtermism from discounting small probabilities. First, I discussed the Low Risks Argument: The probabilities of existential catastrophes are so low that we ought to ignore them. However, even in the next century, the net existential risk and some specific existential risks are above any reasonable discounting thresholds. Naive Discounting faces the Outcome Individuation Problem, so it is unclear what it says. However, an acceptable solution to this problem should not imply that one ought to ignore a net existential risk of 1/6 in the next century. Tail Discounting is more plausible than Naive Discounting. However, as long as there are non-negligible probabilities of better and worse outcomes than a near-term extinction, Tail Discounting will not ignore near-term extinction events even if their associated probabilities are negligible.

The second argument against Longtermism I discussed is the Small Future Argument: Once we ignore very-small-probability scenarios, such as space settlement and digital minds, the expected number of lives in the far future is too small for Longtermism to be true. However, this does not seem true. For example, AI safety leads to Longtermism if humanity's expected lifespan is at least 265 years. Therefore, the Small Future Argument does not undermine Longtermism.

Finally, I discussed the No Difference Argument: The probability that an agent can make a difference to whether or not an existential catastrophe occurs is so small that it should be discounted down to zero. Baseline State Discounting captures this idea naturally. It may also challenge Longtermism, as there is only a tiny probability that Shivani can make a difference to whether or not an existential catastrophe occurs. However, if states are individuated in the most plausible way (i.e., by utilities), Baseline State Discounting violates Statewise Dominance, which makes it less plausible as a theory of instrumental rationality.

Lastly, I argued that Probability Discounting faces Each-We Dilemmas. If Probability Discounting is to avoid Each-We Dilemmas, it needs Collective Difference-Making: Agents must take into account the choices of other people and consider whether the collective can make a difference. However, if we accept Collective Difference-Making, then Probability Discounting does not undermine Longtermism because Shivani and all the other agents together have a non-negligible probability of making a difference.

All in all, I have discussed three ways in which discounting small probabilities might undermine Longtermism. I have argued that these arguments do not succeed. Discounting small probabilities gives no reason to reject Longtermism.

Appendix

A State Discounting and Acyclicity

Earlier I discussed a version of State Discounting on which states are partitioned by comparing prospects to a status quo prospect. But there are different views about how states should be partitioned. On another version of State Discounting, prospects are always compared two at a time, and the possible states of the world are partitioned for every pairwise comparison separately. Alternatively, one could compare all available options at once and partition the states for every choice set separately. Let's call these views *Pairwise State Discounting* and *Set-Dependent State Discounting*, respectively.

> **Pairwise State Discounting:** States are partitioned by comparing two prospects at a time.

> **Set-Dependent State Discounting:** States are partitioned by comparing all prospects at once.

The argument against Longtermism from Pairwise and Set-Dependent State Discounting is similar to that from Baseline State Discounting. Recall that by donating $10,000 to the Against Malaria Foundation, Shivani can save two lives in expectation. By donating the same money to AI safety, she can provide a 1 in 10 billion absolute reduction in the probability of an AI-driven catastrophe in the next 100 years. Instead of partitioning the states by comparing AI safety and the Against

Malaria Foundation to a status quo prospect, Pairwise and Set-Dependent State Discounting partition the states by comparing these two options. Consequently, it ignores tiny differences between these prospects. As the whole choice set only includes two alternatives, both views treat the case similarly.

As before, in order to capture the idea of the No Difference Argument, states must be individuated based on whether or not Shivani makes a difference to an AI-driven catastrophe as follows: In state 1, an AI causes an existential catastrophe no matter what Shivani does. In state 2, an AI does not cause an existential catastrophe if she donates to AI safety, but it will cause an existential catastrophe if she donates to the Against Malaria Foundation. Lastly, in state 3, an AI does not cause an existential catastrophe no matter what she does. Donating to the Against Malaria Foundation saves two lives in all states. Shivani's choice situation is shown in table 10.

Table 10

AI Safety vs. AMF

	State 1	**State 2**	**State 3**
	$p \approx 0.001$	$p = 10^{-10}$	$p \approx 0.999$
AI safety	AI doom	No AI doom	No AI doom
AMF	AI doom + 2 lives	AI doom + 2 lives	No AI doom + 2 lives

As before, if Shivani's discounting threshold is higher than 1 in 10 billion, she ought to ignore the possibility of state 2 obtaining. Consequently, donating to the Against Malaria Foundation is better because it gives a better outcome in states

1 and 3. So, like Baseline State Discounting, Pairwise and Set-Dependent State Discounting challenge Longtermism if they partition states as in table 10.

However, partitioning states as in table 10 leads to a violation of the following principle:

> **Acyclicity:** If $X_1 \succ X_2 \succ \cdots \succ X_n$, then it is not the case that $X_n \succ X_1$.

According to Pairwise and Set-Dependent State Discounting, states might be partitioned differently depending on what other options are available, and this can generate cycles. The former violates Acyclicity within choice sets, while the latter violates Acyclicity across choice sets when two options are compared at a time.

Suppose that Shivani gives a 5% probability for an AI-driven catastrophe and that (implausibly) her discounting threshold is 2%. Next, to see why Pairwise and Set-Dependent State Discounting violate Acyclicity, consider the following options:

> **Acyclicity Violation:**
>
> *Against Malaria Foundation* Saves two lives and gives a 5% probability of an AI-driven catastrophe.
>
> *Pure AI safety* Decreases the probability of an AI-driven catastrophe to 3%.
>
> *Mixed AI safety* Decreases the probability of an AI-driven catastrophe to 4% and, in addition, saves one life in the near-term future.

First, let's compare Pure AI safety to donating to the Against Malaria Foundation (see table 11). States are partitioned in the same way as in table 10. But, in this case, the probability of state 2 is not below the discounting threshold, so there is a non-negligible chance that Shivani can influence whether an AI-driven catastrophe occurs. Consequently, when state 2 is not ignored, donating to Pure AI safety is better than donating to the Against Malaria Foundation.

Table 11
Pure AI Safety Is Better than AMF

	State 1	**State 2**	**State 3**
p	0.03	0.02	0.95
Pure AI safety	Doom	No doom	No doom
AMF	Doom + 2 lives	Doom + 2 lives	No doom + 2 lives

Next, let's compare the Against Malaria Foundation to Mixed AI safety (see table 12). Again, states are partitioned in the same way as in table 10. This time, the probability of state 2* is below the discounting threshold, so Shivani should ignore the possibility of influencing an AI-driven catastrophe. Moreover, when state 2* is ignored, the Against Malaria Foundation is better than Mixed AI safety because it gives a better outcome in states 1* and 3* (two lives saved instead of one). So, now we have that Pure AI safety is better than the Against Malaria Foundation, which is better than Mixed AI safety. It follows by Acyclicity that Mixed AI safety is not better than Pure AI safety.

Table 12

AMF Is Better than Mixed AI Safety

	State 1**	**State 2****	**State 3****
p	0.04	0.01	0.95
Mixed AI safety	Doom + 1 life	No doom + 1 life	No doom + 1 life
AMF	Doom + 2 lives	Doom + 2 lives	No doom + 2 lives

However, when we compare Pure AI safety and Mixed AI safety pair-wise, we find the opposite: Mixed AI safety is better than Pure AI safety (see table 13). As before, states are partitioned the same way as in table 10. In this case, the probability of state 2** is below the discounting threshold, so one should consider Pure AI safety and Mixed AI safety equally effective at reducing the probability of an AI-driven catastrophe. Consequently, Mixed AI safety is better than Pure AI safety because it gives a better outcome in states 1** and 3**. So, now we have that Pure AI safety is better than the Against Malaria Foundation, which is better than Mixed AI safety, which is better than Pure AI safety. This is a violation of Acyclicity.[120]

Table 13

Mixed AI Safety Is Better than Pure AI Safety

	State 1**	**State 2****	**State 3****
p	0.03	0.01	0.96
Pure AI safety	Doom	No doom	No doom
Mixed AI safety	Doom + 1 life	Doom + 1 life	No doom + 1 life

[120] Pairwise and Set-Dependent State Discounting also violate (first-order) Stochastic Dominance. See §4 in Chapter 4 of this thesis.

Pairwise State Discounting violates Acyclicity even within a single choice set; when all three options are available, there is no clear winner. Therefore, it is unclear what Pairwise State Discounting implies and what one ought to choose. Set-Dependent State Discounting, in turn, violates Pair-Wise Acyclicity, that is, it violates Acyclicity when we only compare two options at a time (as in tables 11, 12 and 13). However, if we compare all three options at once, then Set-Dependent State Discounting avoids this violation of Acyclicity, at least if states are partitioned as in table 14. In this case, states 2*** and 3*** have probabilities below the discounting threshold, so Shivani should ignore the possibilities of states 2*** and 3***. Once she does that, the Against Malaria Foundation comes out as the best option because it gives the best outcome in states 1*** and 4***. Within this choice-set, the Against Malaria Foundation is better than Mixed AI safety, which is better than Pure AI safety, which is worse than the Against Malaria Foundation. So, there is no violation of Acyclicity.

Table 14

No Violation of Acyclicity

	State 1* **	**State 2*** **	**State 3*** **	**State 4*** **
p	0.03	0.01	0.01	0.95
Pure AI safety	Doom	No doom	No doom	No doom
Mixed AI safety	Doom + 1 life	Doom + 1 life	No doom + 1 life	No doom + 1 life
AMF	Doom + 2 lives	Doom + 2 lives	Doom + 2 lives	No doom + 2 lives

Note that this case also shows that Set-Dependent State Discounting violates the following intuitively plausible principles:[121]

[121] Sen (1977, pp. 63–66). Contraction Consistency implies Acyclicity. See Sen (1977, p. 67).

> **Contraction Consistency:** For all prospects X and Y, if it is permissible to choose X from the set $\{X, \dots, Y\}$, then it is permissible to choose X from any subset of the set $\{X, \dots, Y\}$.

> **Strong Expansion Consistency:** For all prospects X, Y and Z, if it is permissible to choose X from the set $\{X, \dots, Y\}$, then if it is permissible to choose Y from the set $\{X, \dots, Y, \dots, Z\}$, it is permissible to choose X from the set $\{X, \dots, Y, \dots, Z\}$.

Set-Dependent State Discounting violates Contraction Consistency because it is permissible to choose the Against Malaria Foundation when both AI safety options are available (table 14), but it is not permissible to choose it when only Pure AI safety is available (table 11). On the other hand, Set-Dependent State Discounting violates Strong Expansion Consistency because it is permissible to choose Pure AI safety when the Against Malaria Foundation is the only alternative (table 11). However, it is not permissible to choose Pure AI safety when all three options are available, but it is anyhow permissible to choose the Against Malaria Foundation (table 14).

Set-Dependent State Discounting is choice-set dependent. It implies that what Shivani ought to do depends on what other options are available to her, even if she will not choose them. Consequently, whether or not Longtermism is true in Shivani's situation may also be choice-set dependent. Longtermism may be true if Shivani only considers donating to AI safety and the Against Malaria Foundation. However, if she also considers donating to, for example, asteroid detection, then Longtermism may no longer be true in her situation. In that case, there might be more states of nature because of a greater number of available options. Consequently, the difference-making state(s) might now have probabilities below the discounting threshold. This seems implausible. Having more longtermist options should not make Longtermism harder to achieve. However, one implication of Set-Dependent State Discounting is that adding more options can decrease the probabilities of the difference-making state(s) sufficiently to render Longtermism false.

To summarize, like Baseline State Discounting, the two alternative versions of State Discounting present a challenge to Longtermism. However, they give cyclic recommendations, which makes them less plausible as theories of instrumental rationality.

References

Adams, F. C. (2008), Long-term astrophysical processes, *in* N. Bostrom and M. Cirkovic, eds, 'Global Catastrophic Risks', Oxford University Press, Oxford.

Balfour, D. (2021), 'Pascal's Mugger strikes again', *Utilitas* **33**(1), 118–124.

Beckstead, N. (2013), On the overwhelming importance of shaping the far future, PhD thesis, Rutgers, the State University of New Jersey.

Beckstead, N. and Thomas, T. (2020), 'A paradox for tiny probabilities and enormous values'. Global Priorities Institute Working Paper No.10.

URL: *https://globalprioritiesinstitute.org/nick-beckstead-and-teruji-thomas-a-paradox-for-tiny-probabilities-and-enormous-values/*

Bostrom, N. (2003), 'Astronomical waste: The opportunity cost of delayed technological development', *Utilitas* **15**(3), 308–314.

Bostrom, N. (2009), 'Pascal's Mugging', *Analysis* **69**(3), 443–445.

Bostrom, N. (2013), 'Existential risk prevention as global priority', *Global Policy* **4**(1), 15–31.

Bricker, D. and Ibbitson, J. (2019), *Empty Planet: The Shock of Global Population Decline*, Crown, New York.

Dietz, A. (2016), 'What we together ought to do', *Ethics* **126**(4), 955–982.

Drake, N. (2020), 'Why NASA plans to slam a spacecraft into an asteroid'.
URL: *https://www.nationalgeographic.com/science/article/giant-asteroid-nasa-dart-deflection*

Francis, T. and Kosonen, P. (n.d.), 'Ignore outlandish possibilities'. Unpublished manuscript.

FTX Future Fund (2022), 'Principles'.
URL: *https://ftxfuturefund.org/principles/*

GiveWell (2020), 'GiveWell's cost-effectiveness analyses'.
URL: *https://www.givewell.org/how-we-work/our-criteria/cost-effectiveness/cost-effectiveness-models*

Goodsell, Z. (2021), 'A St Petersburg Paradox for risky welfare aggregation', *Analysis* **81**(3), 420–426.

GPT-3 (n.d.), 'Is artificial general intelligence an existential threat?'.
URL: *https://beta.openai.com/playground*

Grace, K., Salvatier, J., Dafoe, A., Zhang, B. and Evans, O. (2018), 'When will AI exceed human performance? Evidence from AI experts', *Journal of Artificial Intelligence Research* **62**, 729–754.

Greaves, H. and MacAskill, W. (2021), 'The case for strong longtermism'. Global Priorities Institute Working Paper 5–2021.
URL: *https://globalprioritiesinstitute.org/hilary-greaves-william-macaskill-the-case-for-strong-longtermism-2/*

Gustafsson, J. and Kosonen, P. (n.d.), 'Prudential Longtermism'. Unpublished manuscript.

Hájek, A. (2014), 'Unexpected expectations', *Mind* **123**(490), 533–567.

Hey, J. D., Neugebauer, T. M. and Pasca, C. M. (2010), Georges-Louis Leclerc de Buffon's 'Essays on moral arithmetic', *in* A. Sadrieh and A. Ockenfels, eds, 'The Selten School of Behavioral Economics: A Collection of Essays in Honor of Reinhard Selten', Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 245–282.

Isaacs, Y. (2016), 'Probabilities cannot be rationally neglected', *Mind* **125**(499), 759–762.

Kagan, S. (2011), ‘Do I make a difference?’, *Philosophy and Public Affairs* **39**(2), 105–141.

Kreps, D. M. (1988), *Notes on the Theory of Choice*, Westview Press, Boulder.

Kutz, C. (2000), *Complicity: Ethics and Law for a Collective Age*, Cambridge University Press, Cambridge.

Lubin, P. and Cohen, A. N. (n.d.), ‘Don’t forget to look up’. Unpublished manuscript.
URL: *https://arxiv.org/pdf/2201.10663.pdf*

Luce, R. D. and Raiffa, H. (1957), *Games and Decisions: Introduction and Critical Survey*, Wiley, New York.

Lundgren, B. and Stefánsson, H. O. (2020), ‘Against the De Minimis principle’, *Risk Analysis* **40**(5), 908–914.

MacAskill, W. (2019), ‘Longtermism’, Effective Altruism Forum.
URL: *https://forum.effectivealtruism.org/posts/qZyshHCNkjs3TvSem/longtermism*

MacAskill, W., Vallinder, A., Shulman, C., Österheld, C. and Treutlein, J. (2021), ‘The evidentialist’s wager’, *Journal of Philosophy* **118**(6), 320–342.

McMahan, J. (1981), ‘Problems of population theory’, *Ethics* **92**(1), 96–127.

Millett, P. and Snyder-Beattie, A. (2017), ‘Existential risk and cost-effective biosecurity’, *Health Security* **15**(4), 373–383.

Monton, B. (2019), 'How to avoid maximizing expected utility', *Philosophers' Imprint* **19**(18), 1–24.

Musk, E. (n.d.), 'Mars & beyond: The road to making humanity multiplanetary'. **URL:** *http://www.https://www.spacex.com/human-spaceflight/mars/*

Nefsky, J. (2011), 'Consequentialism and the problem of collective harm: A reply to Kagan', *Philosophy and Public Affairs* **39**(4), 364–395.

Nefsky, J. (2015), Fairness, participation, and the real problem of collective harm, *in* M. Timmons, ed., 'Oxford Studies in Normative Ethics', Vol. 5, Oxford University Press, Oxford, pp. 245–271.

Nefsky, J. (2017), 'How you can help, without making a difference', *Philosophical Studies* **174**(11), 2743–2767.

Newberry, T. (2021), 'How cost-effective are efforts to detect near-Earth-objects?'. Global Priorities Institute Technical Report T1–2021. **URL:** *https://globalprioritiesinstitute.org/how-cost-effective-are-efforts-to-detect-near-earth-objects-toby-newberry-future-of-humanity-institute-university-of-oxford/*

Nover, H. and Hájek, A. (2004), 'Vexing expectations', *Mind* **113**(450), 237–249.

Nozick, R. (1969), Newcomb's problem and two principles of choice, *in* N. Rescher, ed., 'Essays in Honor of Carl G. Hempel: A Tribute on the Occasion of His Sixty-Fifth Birthday', Reidel, Dordrecht, pp. 114–146.

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

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

Parfit, D. (1988), 'What we together do'. Unpublished manuscript.

Pulskamp, R. J. (n.d.), 'Correspondence of Nicolas Bernoulli concerning the St. Petersburg Game'. Unpublished manuscript. Accessed through: https://web.archive.org/.
URL: *http://cerebro.xu.edu/math/Sources/NBernoulli/correspondence_petersburg_ game.pdf*

Rees, M. (2003), *Our Final Hour: A Scientist's Warning: How Terror, Error, and Environmental Disaster Threaten Humankind's Future in This Century—on Earth and Beyond*, Basic Books, New York.

Rodriguez, L. (2021), 'Luisa Rodriguez on why global catastrophes seem unlikely to kill us all'.
URL: *https://80000hours.org/podcast/episodes/luisa-rodriguez-why-global-catastrophes-seem-unlikely-to-kill-us-all/*

Russell, J. S. (2021), 'On two arguments for fanaticism'. Global Priorities Institute Working Paper 17–2021.
URL: *https://globalprioritiesinstitute.org/on-two-arguments-for-fanaticism-jeff-sanford-russell-university-of-southern-california/*

Russell, J. S. and Isaacs, Y. (2021), 'Infinite prospects', *Philosophy and Phenomenological Research* **103**(1), 178–198.

Sagan, C. (1997), *Pale Blue Dot: A Vision of the Human Future in Space*, Ballantine Books, New York.

Savage, L. J. (1951), 'The theory of statistical decision', *Journal of the American Statistical Association* **46**(253), 55–67.

Sen, A. (1977), 'Social choice theory: A re-examination', *Econometrica* **45**(1), 53–88.

Smith, N. J. J. (2014), 'Is evaluative compositionality a requirement of rationality?', *Mind* **123**(490), 457–502.

Smith, N. J. J. (2016), 'Infinite decisions and rationally negligible probabilities', *Mind* **125**(500), 1199–1212.

Snyder-Beattie, A., Ord, T. and Bonsall, M. (2019), 'An upper bound for the background rate of human extinction', *Scientific Reports* **9**(1), 11054.

Thorstad, D. (n.d.), 'Existential risk pessimism and the time of perils'. Unpublished manuscript.

United Nations, D. o. E. and Social Affairs, P. D. (2019), 'World population prospects 2019: Highlights'.
URL: *https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf*

Wilkinson, H. (2022), 'In defence of fanaticism', *Ethics* **132**(2), 445–477.

Yudkowsky, E. (2007), 'Pascal's Mugging: Tiny probabilities of vast utilities'.
URL: *http://www.overcomingbias.com/2007/10/pascals-mugging.html*