Tiny probabilities and the value of the far future

Petra Kosonen (Population Wellbeing Initiative, University of Texas at Austin)

Global Priorities Institute | February 2023

GPI Working Paper No . 1-2023

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

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

Petra Kosonen[†]

January, 2023

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, there seems to be 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, Jean Baccelli, Andreas Mogensen, Jake Nebel, 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.

[†]Population Wellbeing Initiative, University of Texas at Austin. I would be grateful for comments: kosonenpetra@gmail.com

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 faced by agents today, our acts' expected influence on the value of the world is mainly determined by their possible consequences in the far future.

Longtermism follows naturally from additive views of value, such as total utilitarianism, combined with a risk-neutral decision theory. 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. 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 we might beneficially influence the far future is by mitigating existential risks.[3] 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.[4]

But, there seems to be 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:[5]

---

[1] MacAskill (2019) and Greaves and MacAskill (2021). See also Bostrom (2003), Beckstead (2013) and Ord (2020).

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

[3] 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.

[4] Bostrom (2013).

[5] Bostrom (2009). This case is based on informal discussions by various people, including

> **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 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 value of paying the mugger is positive. 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 value, the mugger can increase the payoff until the offer has positive expected value. Consequently, expected value maximization recommends that Pascal pay the mugger—and thus, it gives the intuitively wrong recommendation.

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*.[6] 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.

However, this paper argues that Probability Discounting does not undermine Longtermism. Three arguments against Longtermism from Probability Discounting will be discussed. §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

---

Eliezer Yudkowsky (2007). See also Balfour (2021).

[6]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), Lundgren and Stefánsson (2020) and Kosonen (2022, Ch 4 and Ch 5) for criticisms of discounting small probabilities.

an individual makes a difference to whether an existential catastrophe occurs is so small that it should be ignored. But, before going into these arguments, I will first say more about Probability Discounting. This paper focuses on three versions of Probability Discounting: Naive Discounting, Tail Discounting and State Discounting. Next, I will introduce Naive Discounting.

1 Discounting small probabilities

Probability Discounting was originally proposed by Nicolaus Bernoulli.[7] 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."[8] 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?"[9]

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 was the probability of a 56-year-old man dying in one day—an outcome reasonable people usually ignore.[10] Condorcet had a similar justification.[11]

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.[12] He would consider a threshold of $1/2$ irrational and some astronomically small threshold

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

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

[9] Hey et al. (2010, p. 256).

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

[11] See Monton (2019, pp. 16–17).

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

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.[13]

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.[14] Ignoring such outcomes can be done by conditionalizing on the supposition that an outcome of non-negligible probability occurs.[15] 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:

> **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.

---

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

[14] Alternatively, one might have a threshold probability t such that outcomes whose probabilities are at most as great as this threshold are ignored.

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

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 existential risks have probabilities above any reasonable discounting thresholds even in the next century. Naive Discounting faces a problem with individuating outcomes, so it is unclear what it says. And it also violates dominance. Another version of Probability Discounting discussed in this section, namely, *Tail Discounting*, is more plausible, as it solves these problems. 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. For example, 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,[16] between 1 in 600,000 and 1 in 50 from a pandemic,[17] 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.[18]

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 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 4$^{\text{th}}$ of January 2055 at 13:00–14:00', 'extinction due to an asteroid impact on the 4$^{\text{th}}$ 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.

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

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

[18]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).

More generally, Naive Discounting faces the following problem:[19]

> **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.[20] Consequently, Naive Discounting does not undermine Longtermism, at least in this way. However, these relatively high estimates of existential risks can be questioned. Might we, after all, have a challenge to Longtermism?

2.2 Tail Discounting

In addition to the Outcome Individuation Problem, Naive Discounting also faces other problems. For example, it violates dominance.[21] 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.[22] Call the outcomes that fall in the middle of the distribution of possible outcomes 'normal outcomes'. Then, more formally, Tail Discounting states the following:

> **Tail Discounting:** For all prospects X and Y, $X \succsim Y$ if and only if
>
> - $EU(X)_{pd} > EU(Y)_{pd}$ or

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

[20] One possible solution is to individuate outcomes by their values. 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 the net extinction risk is high.

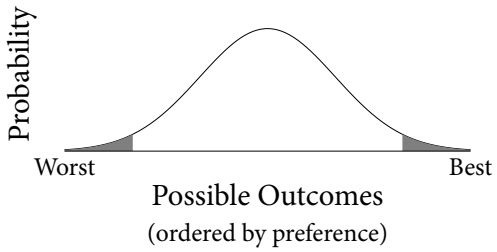
[21] See Isaacs (2016), Smith (2016), Monton (2019, pp. 20–21), Lundgren and Stefánsson (2020, pp. 912–914), Beckstead and Thomas (2020, §2.3) and Kosonen (2022, Ch 4) on Probability Discounting and dominance violations.

[22] Tail Discounting is from Beckstead and Thomas (2020, 2.3).

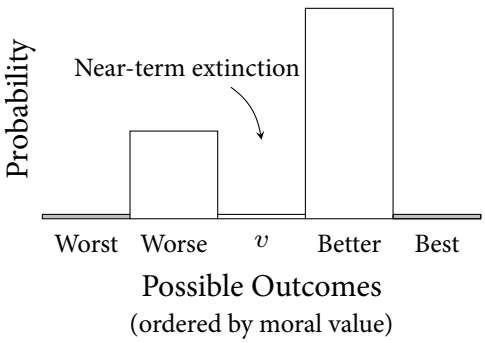
- $EU(X)_{pd} = EU(Y)_{pd}$ and $EU(X) \geq EU(Y)$,

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

Suppose the possible outcomes of some prospect are normally distributed when ordered from the least to the most preferred. Then, Tail Discounting tells us to ignore the grey areas under the curve:



What does Tail Discounting say about extinction risks? Suppose the moral value of a near-term extinction is v. If there are non-negligible probabilities of worse and better outcomes than a near-term extinction, then v falls in the middle of the distribution of possible outcomes' values. Consequently, Tail Discounting will not ignore the possibility of a near-term extinction. 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. And the same is also true for obtaining an

outcome that is at most as good as a near-term extinction. Consequently, Tail Discounting recommends against ignoring the possibility of a near-term extinction.

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.

To summarize, I have discussed the Low Risks Argument: Existential catastrophes are so unlikely that we should 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. But, 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 from Probability Discounting.

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.[23] 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 causes. The cost-effectiveness of antimalarial bednet distribution may be used as an upper bound to attainable near-term benefits per unit of spending.[24] The distribution of insecticide-treated bednets in malarial regions saves a life on average for a little over $4000.[25] Suppose Shivani is thinking how to improve the world the most with her $10,000.[26] Then, by donating to the Against Malaria Foundation,

[23] More precisely, it is not the number of individuals but the amount of value that matters.

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

[25] GiveWell (2020).

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

she can save in expectation 2.5 lives. Therefore, let's 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.[27]

Asteroids. An example of existential risk mitigation that longtermists might focus on is the detection and potential deflection of asteroids.[28] 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.[29] But further work on asteroids is expected to have lower cost-effectiveness.[30] It is estimated that a 10 km+ asteroid has at least a 1% chance of causing human extinction if it collides with the Earth.[31] 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.[32] 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.[33] 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.[34] 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

[27] 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. So, 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, such as generating additional lives in the far future.

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

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

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

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

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

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

[34] Newberry (2021, pp. 5–6).

in 2.4 million.[35] 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.[36]

Pandemics. Another possible cause area longtermists might focus on is the prevention of extinction-level pandemics.[37] 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.[38] 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.[39] 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%.[40] 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.[41]

Artificial Intelligence. Lastly, another possible longtermist cause area is the prevention of an existential catastrophe due to artificial general intelligence.[42] 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%.[43] The same experts gave a 50%

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

[36] $1/(15 \cdot 10^9) \cdot 0.05 \cdot 10000/(1.2 \cdot 10^9) \approx 3 \cdot 10^{-17}$ (1 in 33,000 trillion).

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

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

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

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

[41] $0.01 \cdot 10000/(250 \cdot 10^9) \approx 4 \cdot 10^{-10}$ (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).

[42] See for example Greaves and MacAskill (2021, pp. 14–15).

[43] 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).

chance for high-level machine intelligence occurring by 2061.[44] 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.[45] Furthermore, it is plausible that $1 billion spent on AI safety would decrease the probability of such an outcome by at least 1%.[46] Consequently, $1 billion would provide at least a 0.001% absolute reduction in existential risk.[47] And, 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.[48]

Shivani's options are as follows:

Shivani:

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.

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.

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

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

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

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

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

[48] $0.01 \cdot 10000/10^9 = 10^{-7}$ (1 in 10 million). $0.001 \cdot 0.01 \cdot 10000/10^9 = 10^{-10}$ (1 in 10 billion).

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.[49] 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.[50] 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.[51] Is the expected number of lives in the far future large enough for Longtermism to be true in Shivani's situation (ignoring small probability outcomes)?

Table 2
Expected Population Sizes Required for Longtermism
(With $10,000)

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

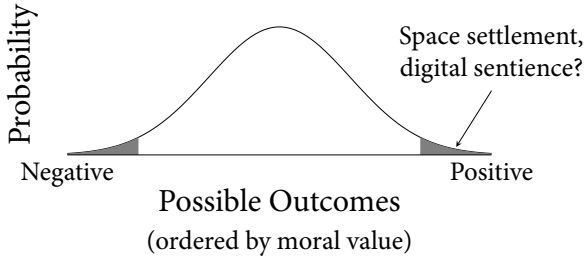
3.2 Is the size of the future large enough?

The truth of 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 risk 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).

[49] $8.3 \cdot 10^{16} \cdot 3 \cdot 10^{-17} \approx 2.5$.

[50] $1.25 \cdot 10^{14} \cdot 2 \cdot 10^{-14} = 2.5$.

[51] $2.5 \cdot 10^{10} \cdot 10^{-10} = 2.5$.



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.[52] This is because humanity might survive 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.[53] 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.[54]

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

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

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

[54] $11 \cdot 10^9 \cdot 87000/80 \approx 1.2 \cdot 10^{13}$ 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).

Longtermism (given that 140 trillion is greater than the required 125 trillion expected future lives).

However, humans are an atypical species, so extinction risk due to natural 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.[55] 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.[56] 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).[57] 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.[58]

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

[55] Adams (2008).

[56] $11 \cdot 10^9 \cdot 10^9/80 \approx 1.4 \cdot 10^{17}$.

[57] $11 \cdot 10^9 \cdot 604 \cdot 10^6/80 > 8.3 \cdot 10^{16}$.

[58] $11 \cdot 10^9 \cdot 909091/80 > 1.25 \cdot 10^{14}$.

billion and a human lifespan of 80 years).[59] 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%.[60] Assuming a constant risk throughout the next 100 years, Ord's (2020, p. 167) estimate for existential risk is below this.[61] So, even if the probability of human extinction was 1/6 in the next 100 years, this would still be low enough for AI safety to lead to 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 we have not taken into account. First, it seems plausible that the risk of extinction per year is not constant. 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,

[59] $11 \cdot 10^9 \cdot 182/80 > 2.5 \cdot 10^{10}$.

[60] $1/0.00377 \approx 265$. 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).

[61] Existential risk in the next 100 years is 1/6 if the risk per year is 0.18%. 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).

after which the yearly risk of extinction is significantly lower.[62] If we now live in a 'time of perils' after which the yearly risk of extinction is significantly lower, existential risk mitigation more easily leads to Longtermism.[63]

To summarize, 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. Thus, the Small Future Argument does not undermine Longtermism. 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. 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.

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.

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*

[62] 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. 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.

[63] 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."

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. 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.[64] But which version of Probability Discounting allows her to do this?

Table 3
Absolute Reductions of Extinction Risk 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 proba-

[64] 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.

bilities instead—let's call this *State Discounting*.[65]

> **State 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 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 paper 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.[66] Let's call this view *Baseline State Discounting*.

---

[65] 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.

[66] 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. These two views violate Acyclicity, while Baseline State Discounting violates Statewise Dominance. See Kosonen (2022, Ch 4 and Ch 6).

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

AMF	Two lives saved
Do nothing	No 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.

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, the next section presents a response to the No Difference Argument.[67]

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

[67] See Kosonen (2022, Ch 4 and Ch 6) for criticism of Baseline State Discounting.

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."[68] To see how Baseline State Discounting (and Probability Discounting more generally) faces Each-We Dilemmas, consider the following case (see table 6):[69]

> **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.

Table 6
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, Baseline State Discounting recommends against attempting to

[68] 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.

[69] One can construct similar Each-We Dilemmas against versions of Probability Discounting that ignore very-small-probability outcomes.

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, 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.[70] 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.[71] The expected benefits of voting can nonetheless be great because if one's vote makes a difference, it will impact millions of people.[72] 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 Baseline State Discounting (and Probability Discounting in general) 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.[73]

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

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

[72] Parfit (1984, pp. 73–75).

[73] 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. 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.

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. 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.[74] This estimate is 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.

[74] $0.001 \cdot 0.01 = 0.00001$. Greaves and MacAskill (2021, pp. 14–15) estimate that there is at 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).

The details of Collective Difference-Making do not matter for the purposes of this paper, so I will only briefly mention some possible justifications for and problems with it. 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, which makes it implausible as well. Either way, Probability Discounting does not undermine Longtermism. Next, I will discuss how Collective Difference-Making could be justified.

5.2 Justifications for Collective Difference-Making

Collective Reasons. 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 so, one could be part of a group of people who together could make a difference.[75] For example, some argue that we have collective reasons for action.[76] On this view, groups, like individuals, have reasons to make outcomes better, benefit other people, avoid harming other people and benefit themselves. 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.[77] 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

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

[76] 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).

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

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.

Causation Without a Difference. Others, in turn, argue that one's act can be part of causing some outcome without making a difference.[78] This can happen when the outcome not happening would be at least partly a result of there not having been enough similar acts.[79] 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.[80]

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

Rule-Consequentialism. Collective Difference-Making can also be justified with, for example, rule-consequentialism. Rule-consequentialism states that

[78] Nefsky (2017).

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

[80] Nefsky (2017, p. 2758).

[81] 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."

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.

Evidential Decision Theory. 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.[82]

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.[83] Evidential Decision Theory recommends 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. Similarly, Evidential Decision Theory recommends Shivani to mitigate existential risks if doing so provides sufficient evidence of others mitigating these risks as well. However, Evidential Decision Theory does not solve Each-We Dilemmas in cases where one's actions do not provide suitably

[82] See Nozick (1969).

[83] 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 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.

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.

Agents. 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 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.[84] On this view, for example causally disconnected intelligent aliens do not count.

Separability. 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:[85]

---

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

[85] 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).

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 depends on what choices other distant agents face.[86] 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:

> **Asteroid 2:** Asteroids are heading toward different planets (one for each planet), and they will almost certainly hit unless they are stopped.

[86] Wilkinson (2022, §6) shows that denying fanaticism leads to violations of Separability (or first-order Stochastic Dominance). 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).

> 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.[87]

Cluelessness. 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 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.[88]

[87] 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.

[88] 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.

Details. 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 paper. 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, the No Difference Argument does not undermine Longtermism. However, discounting small probabilities might still be relevant to 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.

But, 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, I argued that Baseline State Discounting (and Probability Discounting in general) 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. But, 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.

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.

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

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.

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/*

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.

Kosonen, P. (2022), Tiny Probabilities of Vast Value, PhD thesis, University of Oxford.

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

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.

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.

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/*

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

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*