Tiny Probabilities and the Value of the Far Future

Petra Kosonen

kosonenpetra@gmail.com

The Lancet Commission on 21st Century Global Threats to Health

- My talk is relevant to both Lancet Commission questions:
  1. What is a legitimate timeframe over which the Commission and the users of the Commission findings should be concerned about global threats?
  2. Considering that global threats are highly uncertain, should the Commission adopt a risk neutral stance or a precautionary approach?

Longtermism

- 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.

- 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]

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

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

Foreseeably affecting the far future

- 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.

Existential risks

- 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.

Existential risks

Risks that threaten the destruction of humanity's long-term potential.

[3] Bostrom (2013).

Existential risks

- An individual agent may only have a small impact on the probability of an existential risk.
- However, 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]

[4] Bostrom (2013).

Tiny probabilities of huge value

- 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 in some decision theoretic cases.
- Consider, for example, the following case:[5]

Pascal's Hell

Satan offers Pascal a deal: If a coin lands on heads, he will create a million Graham's number of happy Earth-like planets; otherwise everyone on Earth will suffer excruciating pain for one billion years. The probability of heads is one-in-Graham's-number.

[5] Kosonen (2022, pp. 2-4). This case is inspired by Bostrom's (2009) *Pascal's Mugging*, which is based on informal discussions by multiple people, such as Eliezer Yudkowsky (2007). See also Balfour (2021).

Pascal's Hell

- Should Pascal accept the deal?
- The probability of the great outcome is tiny, so accepting the deal will almost certainly result in a very bad outcome (a billion years of torture for everyone!)
- However, as the great outcome is amazingly good, Pascal is forced to conclude that accepting the deal has positive expected value.

Probability Discounting

- 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*.
- 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.

Outline

- However, I'll argue that Probability Discounting does not undermine Longtermism.
- In the paper, I discuss three arguments against Longtermism from Probability Discounting:
  1. The probabilities of existential catastrophes are so low that one ought to ignore them.
  2. 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.
  3. The probability that an individual makes a difference to whether an existential catastrophe occurs is so small that it should be ignored.
- In this talk, I'll discuss the first and the third arguments.
- However, before going into these arguments, I will first say more about Probability Discounting.

History of Probability Discounting

- Probability Discounting was originally proposed by Nicolaus Bernoulli.
- 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."[6]
- 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?"[7]

[6] Pulskamp (n.d., p. 2).
[7] Hey et al. (2010, p. 256).

Discounting threshold

- Some have suggested possible discounting thresholds.

Buffon (1777)

$t = 1/10,000$

Condorcet (1785)

$t = 1/144,768$

Monton (2019)

$t = 1$ in 2 quadrillion (5×10^{-16})

- Subjective preference

- 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.[8]
- Condorcet had a similar justification.[9]

[8] Hey et al. (2010, p. 257). Photos: Wikimedia Commons.
[9] See Monton (2019, pp. 16–17).

Discounting threshold

- 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.[10]
- 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.

[10] Monton (2019, §6.1).

Small probabilities of *what*?

- 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.
- Ignoring such outcomes can be done by conditionalizing on the supposition that an outcome of non-negligible probability occurs.[11]
- After conditionalization, options are compared by their 'probability-discounted expected utilities'.

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

Naive Discounting

- 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

For all prospects X and Y, X is at least as good as Y if and only if $EU(X)_{pd} \geq EU(Y)_{pd}$, where $EU(X)_{pd}$ and $EU(Y)_{pd}$ are obtained by conditionalizing on the supposition that some outcome of non-negligible probability occurs.

Summary

- 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.

Low Risks Argument

- 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 catastrophes 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.

Existential risks in this century

- 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: "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."[12]

[12] Rees (2003).

Existential risks in this 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.
- 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,[13] between 1 in 600,000 and 1 in 50 from a pandemic,[14] and a 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.[15]

[13] See Ord (2020, p. 71).
[14] Millett and Snyder-Beattie (2017).
[15] Greaves and MacAskill (2021, pp. 14–15).

Existential risks in this century

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.

Forecasters

- In a recent study that asked talented forecasters what they think about existential risks, the median 'superforecaster' predicted a 9% chance of global catastrophe (that kills at least 10%) and a 1% chance of extinction by year 2100.[16]
- Here are some superforecasters' estimates for extinction risks from various causes by year 2100:

1. AI extinction: 0.38%
2. Engineered pathogen extinction: 0.01%
3. Nuclear extinction: 0.074%
4. Total extinction risk: 1%

[16] Karger et al. (2023, p. 4). Karger et al. (2023, p. 4, n.4) define extinction as reduction of the global population to less than 5000.

Individuating outcomes

- 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 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.

Outcome Individuation Problem

- More generally, Naive Discounting faces the following problem:[17]

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?

[17] See also Beckstead and Thomas (forthcoming, p. 13).

Outcome Individuation Problem

- 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 (Toby Ord) or a 1% extinction risk (forecasters).
- Consequently, Naive Discounting does not undermine Longtermism, at least in this way.

Summary

- 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 can ignore a net existential risk of 1/6 in the next century (Toby Ord) or a 1% extinction risk (forecasters).
- To conclude, the Low Risks Argument does not undermine Longtermism.

No Difference Argument

- The second 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.

Probability Discounting and Each-We Dilemmas

- However, next I'll argue 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.

Probability Discounting and Each-We Dilemmas

- According to Parfit, 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."[18]
- To see how Probability Discounting faces Each-We Dilemmas, consider the following case:

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.

[18] Parfit (1984, p. 91).

Probability Discounting and Each-We Dilemmas

- In state 1, the asteroid will hit the Earth no matter what the agent chooses; in state 3, the asteroid won't hit the Earth no matter what the agent chooses; and in state 2, the agent can make a difference to whether or not the asteroid hits the Earth.
- However, the probability of state 2 happening is below the discounting threshold, so the possibility of this state should be ignored.
- Then doing nothing is better than attempting to stop the asteroid because it gives a better outcome in states 1 and 3.

TABLE 2
ASTEROID

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

Probability Discounting and Each-We Dilemmas

- So, versions of Probability Discounting that recommend ignoring tiny probabilities of making a difference would in this case recommend against attempting to stop the asteroid.
- Consequently, the asteroid will almost certainly hit the Earth—which could have been prevented almost certainly had enough agents attempted to do so.

TABLE 3
ASTEROID

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

Collective Difference-Making

- 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.

Collective Difference-Making

- The probability that we together can make a difference to existential risks seems non-negligible.
- For example, Greaves and MacAskill (2021, pp. 14–15) estimate that if we spend $1 billion on AI safety, we can plausibly provide at least a 1 in 100,000 absolute reduction in the probability of an AI-driven catastrophe.[19]
- So, if we should accept Collective Difference-Making, then—plausibly—Probability Discounting does not undermine Longtermism.
- We should not ignore the possibility of making a difference because we together with all the other agents have a non-negligible chance of preventing an existential catastrophe.

[19] 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%.

Collective Reasons

- I will not evaluate the plausibility of Collective Difference-Making in this talk.
- Instead, 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 all the agents together have a non-negligible probability of making a difference.
- Either way, the No Difference Argument does not undermine Longtermism.

Conclusion

- I have discussed two 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 can ignore a net existential risk of 1/6 in the next century (Toby Ord) or a 1% chance of extinction by 2100 (forecasters).

Conclusion

- Next, I discussed the No Difference Argument: The probability that an agent makes a difference to whether or not an existential catastrophe occurs is so small that it should be discounted down to zero.
- This argument may challenge Longtermism, as there is only a tiny probability that we can make a difference to whether or not an existential catastrophe occurs.
- However, I argued that Probability Discounting faces Each-We Dilemmas, and if it 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 we and all the other agents together have a non-negligible probability of making a difference.

Conclusion

- All in all, I have discussed two 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 I

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. (forthcoming), 'A paradox for tiny probabilities and enormous values', *Nous* .

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.

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/*

References II

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.

Karger, E., Rosenberg, J., Jacobs, Z., Hickman, M., Hadshar, R., Gamin, K., Smith, T., Williams, B., McCaslin, T. and Tetlock, P. E. (2023), 'Forecasting existential risks: Evidence from a long-run forecasting tournament'. Forecasting Research Institute Working Paper No. 1. **URL:** *https://static1.squarespace.com/static/635693acf15a3e2a14a56a4a/t/64abffe3f024747dd0e38d71/1*

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

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

References III

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.

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

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

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.

References IV

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

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