High risk, low reward: A challenge to the astronomical value of existential risk mitigation

David Thorstad (Global Priorities Institute, University of Oxford)

Global Priorities Institute | July 2023

GPI Working Paper No . 6-2023

High risk, low reward: A challenge to the astronomical value of existential risk mitigation

Forthcoming in *Philosophy and Public Affairs.*
Penultimate draft. Please cite published version.

Abstract: Many philosophers defend two claims: the *astronomical value thesis* that it is astronomically important to mitigate existential risks to humanity, and *existential risk pessimism*, the claim that humanity faces high levels of existential risk. It is natural to think that existential risk pessimism supports the astronomical value thesis. In this paper, I argue that precisely the opposite is true. Across a range of assumptions, existential risk pessimism significantly reduces the value of existential risk mitigation, so much so that pessimism threatens to falsify the astronomical value thesis. I argue that the best way to reconcile existential risk pessimism with the astronomical value thesis relies on a questionable empirical assumption. I conclude by drawing out philosophical implications of this discussion, including a transformed understanding of the demandingness objection to consequentialism, reduced prospects for ethical longtermism, and a diminished moral importance of existential risk mitigation.

Keywords: Longtermism, existential risk, catastrophic risk, ethics of risk.

1. Introduction

Derek Parfit (1984) invites us to consider two scenarios. In the first, a war kills ninety-nine percent of the world's human population. Such an event, Parfit urges, would be a great tragedy. Billions would die and the rest would suffer terribly. Nations would fall.

Cities and monuments would be destroyed. Recovering from such a catastrophe would take centuries.

In a second scenario, the same war kills every living human. This event, Parfit holds, would be many times worse than the first. Unthinkably many future lives would fail to be lived Greaves and MacAskill 2021). Our projects would remain forever incomplete and our purposes unfulfilled (Bennett 1978; Knutzen forthcoming; Riedener 2021). All nations, cultures and families would end (Scheffler and Kolodny 2013). There would be no more art, science, music or philosophy (Parfit 1984). A human species that might have flourished for billions of years would find itself extinguished in its infancy (Bostrom 2003; Ord 2020).

Some followers of Parfit have drawn the lesson that it is overwhelmingly important to mitigate *existential risks*: risks of existential catastrophes involving "the premature extinction of Earth-originating intelligent life or the permanent and drastic destruction of its potential for desirable future development" (Bostrom 2013, p. 15). For example, we might work to regulate chemical and biological weapons or to reduce the threat of nuclear conflict (Bostrom and Ćirković 2011; MacAskill 2022; Ord 2020). Mitigating existential risk is frequently held to be not only valuable, but also astronomically more valuable than tackling important global challenges such as poverty, inequality, global health or racial injustice (Bostrom 2013; Ord 2020). The reason given is that existential risk mitigation offers a chance of tremendous gain: the continued survival and development of humanity. Given the mind-boggling scale of what might be lost, anything that we can do to prevent existential catastrophe may have astronomical value. Let the *astronomical value thesis* be the claim that the best available options for reducing existential risk today have astronomical value.

The astronomical value thesis is often combined with alarmingly high estimates of current existential risk. Toby Ord puts the risk of existential catastrophe by 2100 at “one in six: Russian roulette” (Ord 2020, p. 46). The Royal Astronomer Martin Rees gives a 50% chance of civilizational collapse by 2100 (Rees 2003). And participants at the Oxford Global Catastrophic Risk Conference in 2008 estimated a median 19% chance of human extinction by 2100 (Sandberg and Bostrom 2008).

Let *existential risk pessimism* be the view that existential risk this century is very high — for concreteness, say twenty percent. It is often supposed that existential risk pessimism bolsters the case for the astronomical value thesis. After all, we should usually do more to address probable threats than to address improbable threats. In this paper, I use a series of models to draw a counterintuitive conclusion. Across a range of assumptions, existential risk pessimism not only fails to increase the value of existential risk mitigation, but in fact substantially decreases it, so much so that existential risk pessimism threatens to falsify the astronomical value thesis (Sections 2-3). I suggest that the best way to reconcile existential risk pessimism with the astronomical value thesis relies on a questionable empirical assumption, the time of perils hypothesis that risk is high now, but will soon fall to a very low level (Sections 4-6). I argue that the time of perils hypothesis is not well supported. If that is right, then existential risk pessimism threatens to tell against the astronomical value thesis. Existential risk mitigation may yet be valuable, but perhaps not astronomically so.

I conclude by drawing out four philosophical consequences of this discussion: a transformed understanding of the demandingness objection to consequentialism (Section 7.1); a challenge to ethical longtermism (Section 7.2); a reduced need for controversial forms of temporal discounting (Section 7.3); and a diminished moral importance of existential risk mitigation (Section 7.4). Proofs are in Appendix A, with additional models in Appendix B.

An important feature of my argument is that it does not rely on ethical or decision-theoretic assumptions which defenders of the astronomical value thesis may be likely to reject. Many recent arguments against the astronomical value thesis have questioned decision-theoretic, consequentialist or population-ethical assumptions used to motivate it ([removed x3]; Lloyd 2021; Mogensen 2022). My argument does not rely on any such maneuvers. The argument in this paper is compatible with standard versions of expected utility theory, interpreted in consequentialist fashion along a range of population axiologies including totalism. In doing this, my aim is to meet the pessimist on her own turf in order to build a case against the astronomical value thesis that may be persuasive even to the pessimist herself.[1]

2. The Simple Model

In this section, I present a Simple Model of existential risk mitigation due to Toby Ord (2020, [removed]). On this model, it will turn out that existential risk pessimism has no effect on the value of existential risk mitigation, and also that the astronomical value thesis is false. Section 3 then considers how the astronomical value thesis fares across modifications of the Simple Model.

[1] However, these arguments would strengthen the conclusions of this paper by further reducing the axiological or deontic importance of existential risk mitigation. In that sense, they might be viewed as important complements to the present project.

The Simple Model makes three assumptions. First, it assumes that each century of human existence has some constant value v. Second, it assumes that humans face a constant level of per-century existential risk r. And third, it assumes that all existential risks are risks of human extinction, so that no value will be realized after an existential catastrophe. These are restrictive assumptions, and Section 3 will consider what happens when we relax them.[2] But under these assumptions, we can evaluate the expected value of the current world W, incorporating possible future continuations, as follows:

(Simple Model) $V[W] = v\sum_{i=1}^{\infty}(1-r)^i = v(1-r)/r$.

On this model, the value of our world today depends on the value v of a century of human existence as well as the risk r of existential catastrophe. Setting r at a pessimistic 20% values the world at a mere four times the value of a century of human life, whereas an optimistic risk of 0.1% values the world at the value of nearly a thousand centuries.

Now suppose that you can act to reduce existential risk in your own century. More concretely, you can take some action X which will reduce risk this century by some fraction f, from r to $(1-f)r$. However, let us suppose that your actions will have no effect on future risks. What is the value of your action?

On the Simple Model, it turns out that $V[X] = fv$. This result is surprising for two reasons. First, the value of action X is entirely independent of the current level of existential risk. Halving existential risk from 20% to 10% has the same value as halving it

[2] I will not explicitly consider the consequences of distinguishing between extinction and non-extinction catastrophes. On many natural ways of relaxing this assumption, we may recover interesting normative consequences, but the heightened importance of existential risk reduction will not be among them.

from 2% to 1%. This means that the truth or falsity of existential risk pessimism is entirely irrelevant to the value of existential risk mitigation. By contrast, we might have thought that existential risk pessimism increases the value of existential risk mitigation.

A second surprising result is that the Simple Model does not support the astronomical value thesis. Although the future itself may be astronomically valuable, the expected value of reducing existential risk in this century is capped at the value v of an additional century of human existence. This means that interventions which present a small chance of preventing existential catastrophe in this century may not be obviously more valuable than other altruistic interventions, such as work done to mitigate extreme poverty. By way of example, an action which reduces the risk of existential catastrophe in this century by one trillionth would have, in expectation, one trillionth as much value as a century of human existence. Lifting several people out of poverty from among the billions who will be alive in this century may be more valuable than this. In this way, the Simple Model presents a *prima facie* challenge to the astronomical value of existential risk mitigation.

In this section, we developed a Simple Model of existential risk reduction. We saw that on this model, existential risk pessimism has no bearing on the value of existential risk mitigation and the astronomical value thesis is false. Can the pessimist increase the value of existential risk mitigation by modifying the Simple Model? In the next section, I consider four ways that the pessimist might proceed.

3. Modifying the Simple Model

In this section, I extend an analysis by Ord to consider four ways in which the Simple Model may be modified. I argue that the last of these strategies is the most viable. This strategy involves introducing an empirical hypothesis, the time of perils hypothesis, which will be evaluated in Sections 4-6.

3.1 Absolute versus relative risk reduction

In working through the Simple Model, we considered the value of reducing existential risk by some fraction f of its original amount. But this might seem like comparing apples to oranges. Reducing existential risk from 20% to 10% may be more difficult than reducing existential risk from 2% to 1%, even though both involve reducing existential risk to half of its original amount. Wouldn't it be more realistic to compare the value of reducing existential risk from 20% to 19% with the value of reducing risk from 2% to 1%?

More formally, we were concerned about *relative reduction* of existential risk from its original level r by the fraction f, to $(1-f)r$. Instead, the objection goes, we should have been concerned with the value of *absolute risk reduction* from r to $r-f$. Will this change help the pessimist?

It will not. On the Simple Model, the value of absolute risk reduction is fv/r. Now the value of risk reduction is no longer independent of the current level of risk r. Rather, we have made matters worse for the pessimist: the value of risk reduction *decreases* the more pessimistic we are about current existential risk. Multiplying the level of current risk r by some fixed amount N reduces the value of absolute risk reduction by N, so that for example absolute risk reduction is a hundred times more valuable if we estimate risk at

0.2% rather than 20%. Here pessimism serves to lower, rather than raise the value of existential risk mitigation. That is not what the pessimist wanted. What else might the pessimist do to support the astronomical value thesis?

3.2 Value growth

The Simple Model assumed that each additional century of human existence has some constant value v. However, on many population axiologies the value of an additional century of human existence is likely to increase over time. That is because future centuries may support larger populations and may support these populations at higher levels of welfare and with longer lifespans. What happens if we modify the Simple Model to account for value growth?

In this section, we will see that accounting for value growth does boost the case for existential risk mitigation across the board, but that on its own value growth is unlikely to ground the astronomical value thesis. We will also see that as increasingly optimistic assumptions about value growth are considered, pessimism looms larger as a roadblock to the astronomical value thesis. Appendix B extends the discussion in this section to cover highly optimistic growth assumptions.

Let v be the value of the present century. We might assume that value grows linearly over time, so that the value of the N_{th} century from now will be N times as great as the value of the present century, if we live to reach it.

(Linear Growth) $V[W] = v \sum_{i=1}^{\infty} i\,(1-r)^{i} = v(1-r)/r^{2}$.

On this model, the value of reducing existential risk by some (relative) fraction f is fv/r. Somewhat generously, we might also consider an optimistic growth model in which value grows quadratically over time, so that the N_{th} century will be N^2 times as valuable as the present century.

(Quadratic Growth) $V[W] = v\sum_{i=1}^{\infty} i^2 (1-r)^i = v(1-r)(2-r)/r^3$.

On this model, the value of reducing existential risk by f is $fv(2-r)/r^2$. These models have two noteworthy consequences.

First, the value of relative risk reduction remains capped at a modest v/r on the linear growth model and a somewhat more generous $2v/r^2$ on the quadratic growth model. Table 1 illustrates the value of a 10% reduction in existential risk this century under a variety of views about per-century risk. Under linear growth, even optimistic views about per-century risk assign relatively modest value to existential risk reduction. By contrast, quadratic growth opens the possibility for risk reduction to carry astronomical value. But this is only possible if we abandon pessimism about existential risk. Even under quadratic growth, if we adopt a pessimistic 20% estimate of per-century risk, then reducing relative risk this century by ten percent produces in expectation less than five times the value of the current century. This means that pessimists will have trouble grounding the astronomical value thesis, even under optimistic growth models.

Table 1: Value of 10% relative risk reduction across growth models and risk levels

	r = 0.2	**r = 0.02**	**r = 0.002**	**r = 0.0002**
Linear growth	0.5v	5v	50v	500v
Quadratic growth	4.5v	495v	49,950v	5*10^6v

Second, as we adopt increasingly optimistic growth assumptions, existential risk pessimism ever more strongly devalues existential risk mitigation. Under linear growth, the value of existential risk mitigation varies inversely with per-century risk r, so that adopting a pessimistic 20% estimate of existential risk devalues existential risk by a hundredfold by comparison with an optimistic 0.2% estimate of existential risk. But under quadratic growth, the value of existential risk mitigation varies inversely with the square of r, so that a pessimistic 20% estimate devalues risk-mitigation by a factor of almost 10,000 compared to an optimistic 0.2% risk estimate. Here we begin to gain stronger evidence that pessimism itself is among the primary roadblocks for the astronomical value thesis.

In this section, we saw that considering value growth will increase the value of existential risk mitigation, but that the boost will be modest unless we also weaken our pessimism about existential risk. We also saw that increasingly optimistic assumptions about growth strengthen the tension between existential risk pessimism and the astronomical value thesis. What else might the pessimist do to ground the astronomical value thesis?

3.3 Global risk reduction

The Simple Model assumes that we can only affect existential risk in our own century. This may seem implausible. Our actions affect the future in many ways. Why couldn't our actions reduce future risks as well?

Now it is not implausible to assume that our actions could have measurable effects on existential risk in nearby centuries. Perhaps we can found international institutions dedicated to the prevention of existential risk, and perhaps these institutions will stand for several centuries. But this will not be enough to save the pessimist. On the Simple Model, cutting risk over the next N centuries all the way to zero confers only N times the value of the present century, which is not significantly more than the value of cutting risk in the present century. To salvage the astronomical value thesis, we would need to imagine that our actions today can significantly alter levels of existential risk across very distant centuries. That is less plausible. Are we to imagine that institutions founded to combat risk today will stand or spawn descendants millions of years hence?

More surprisingly, even if we assume that actions today can significantly lower existential risk across all future centuries, this assumption may still not be enough to ground the astronomical value thesis. Consider an action X which reduces per-century risk by the fraction f in all centuries, from r to $(1-f)r$ each century. On the Simple Model, the value of X is then $\frac{f}{1-f}\frac{v}{r}$. Two features of this result deserve note.

First, unlike in previous sections the value of existential risk reduction is now unbounded in the fraction f by which risk is reduced. Under even the most miserly valuation v of a century of human existence and even the most pessimistic estimate r of per-century risk, a 100% reduction in per-century risk carries infinite value, and more generally we can find fractional reductions f in per-century risk which carry arbitrarily high value. However, although the value of existential risk reduction is in principle unbounded, in practice this value may be modest if we are pessimistic about existential risk.[3] By way of illustration, setting r to a pessimistic 20% values a 10% relative reduction

[3] This model does suggest that if we could act today to eliminate existential risk across all future times, that act would have infinite value. However, I know of no proposed interventions which could do this.

in existential risk across all centuries at once at a modest five-ninths of the value of the present century. Even a 90% reduction in risk across all centuries would carry just forty-five times the value of the present century. Hence even the highly optimistic assumption that we can reduce risk across all centuries at once may not be enough to salvage the astronomical value thesis.

Second, at the risk of rehearsing a tired theme, the value of risk reduction once again varies inversely with the current level r of existential risk. As before, pessimism lowers rather than raises the value of existential risk mitigation. It seems likely that pessimism itself must be tempered in order to salvage the astronomical value thesis. In the next subsection, I consider the most plausible way to do this.[4]

3.4 The time of perils

Pessimists often argue that humanity is living through a uniquely perilous period of our history (Aschenbrenner 2020; Ord 2020; Rees 2003; Sagan 1997). Rapid technological growth has given humanity the means to quickly destroy ourselves. If we learn to manage the risks posed by new technologies, then we will enter a period of relative safety. But until

[4] Pessimists might also hold that human survival would have infinite value on other grounds. This suggestion has been raised before (Tarsney and Wilkinson 2023) and to a large extent falls outside the scope of this paper. Three responses to the possibility of promoting infinite value may be worth exploring. First, it is unclear whether humans can promote infinite value given that there is only a finite amount of space within our future light cone. Second, very small chances of promoting infinite value may strengthen worries about fanaticism (Monton 2019, Smith 2014) in which small probabilities of astronomical gain are given too much weight in decisionmaking. Third, to a large extent infinite decision theory is still up in the air, so it may be prudent to avoid drawing normative conclusions which depend on particular views in infinite decision theory.

we do, we are vulnerable to any number of existential catastrophes that could arise from the misuse of new technologies.

This view is often attributed to the astronomer Carl Sagan, who put the point as follows:

> It might be a familiar progression, transpiring on many worlds … life slowly forms; a kaleidoscopic procession of creatures evolves; intelligence emerges … and then technology is invented. It dawns on them that there are such things as laws of Nature … and that knowledge of these laws can be made both to save and to take lives, both on unprecedented scales. Science, they recognize, grants immense powers. In a flash, they create world-altering contrivances. 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 [who] are not so lucky or so prudent, perish. (Sagan 1997, p. 173).

Following Sagan, let the *time of perils hypothesis* denote the view that existential risk will remain high for several centuries, but drop to a low level if humanity survives this time of perils.[5] Could the time of perils hypothesis salvage the astronomical value of existential risk reduction?

To operationalize the time of perils hypothesis, let N be the length of the *perilous period*: the number of centuries for which humanity will experience high levels of risk.

[5] The time of perils hypothesis is related to the *hinge of history hypothesis* that we are living at an especially influential time of history. For discussion see Parfit (2011) and Mogensen (2019).

Assume we face constant risk r throughout the perilous period, with r set to a pessimistically high level. If we survive the perilous period, existential risk will drop to the level r_l of *post-peril risk*, where r_l is much lower than r.

On this model, the value of the world today is:

(Time of Perils) $V[W] = \sum_{i=1}^{N} v\,(1-r)^i + (1-r)^N \sum_{i=1}^{\infty} v\,(1-r_l)^i$.

That works out to an unwieldy

$$V[W] = (1-(1-r)^N)\frac{1-r}{r}v + (1-r)^N \frac{1-r_l}{r_l}v$$

but with some notation, we can get a good handle on the model.

Let V_{PERIL} = $\sum_{i=1}^{\infty} v\,(1-r)^i$ be the value of living in a world forever stuck at the perilous level of risk and V_{SAFE} = $\sum_{i=1}^{\infty} v\,(1-r_l)^i$ be the value of living in a post-peril world. Let SAFE be the proposition that humanity will reach a post-peril world and note that Pr(SAFE) = $(1-r)^N$. Then the value of the world today is a probability-weighted average of the values of the safe and perilous worlds.

$$V[W] = \Pr(\neg\text{SAFE})\,V_{PERIL} + \Pr(\text{SAFE})\,V_{\text{SAFE}}.$$

As the length N of the perilous period and the perilous risk level r trend upwards, the value of the world tends towards the low value V_{PERIL} of the perilous world envisioned by the simple model. But as the perilous period N shortens and the perilous risk r decreases, the value of the world tends towards the high value V_{SAFE} of a post-peril world. These same trends will reappear when we ask after the value of existential risk reduction.

Let X be an action which reduces existential risk in this century by the fraction f, and assume that the perilous period lasts at least one century. Then we have:

$$V[W] = fv[1-(1-r)^N] + r(1-r)^{N-1}fV_{\text{SAFE}}.$$

This equation decomposes the value of X into two components, corresponding to the expected increase in value (if any) that will be realized during the perilous and post-peril periods. The first term, $fv[1-(1-r)^N]$ is bounded above by v, so will be relatively negligible. The case for existential risk mitigation is therefore primarily driven by the second term, $r(1-r)^{N-1}fV_{\text{SAFE}}$, representing the heightened prospect of surviving the time of perils and realizing value thereafter. Call this the *crucial factor*.

The crucial factor may indeed be high enough to bear out the astronomical value thesis, but only if two conditions are satisfied. First, the perilous period N must be short. Because the crucial factor decays exponentially in N, a long perilous period will tend to make the crucial factor quite small. Second, the post-peril risk r_l must be low. The value V_{SAFE} of a post-peril future is determined entirely by the level of post-peril risk, and we saw in Section 2 that this value cannot be high unless risk is very low.

To see how these conditions play out in practice, assume a pessimistic 20% level of risk during the perilous period. Table 2 illustrates the value of a 10% reduction in relative risk across various assumptions about the length of the perilous period and the level of post-peril risk. With a short two-century perilous period and a low 0.1% level of post-peril risk, this action X is as valuable as 160 centuries of additional human life. Building in value growth, X may well have astronomical value. But as the perilous period lengthens or the post-peril risk increases, the value of X decays quickly to its impact on the immediate century. As the perilous period approaches 50 centuries or the post-peril risk approaches even a modest 1%, it becomes very hard to see how even further modifications of the model could assign very high value to X.

Table 2: Value of 10% relative risk reduction against post-peril risk and perilous period length

	N = 2	**N = 5**	**N = 10**	**N = 20**	**N = 50**
$\mathbf{r_l = 0.01}$	1.6v	0.9v	0.4v	0.1v	0.1v
$\mathbf{r_l = 0.001}$	16.0v	8.3v	2.8v	0.4v	0.1v
$\mathbf{r_l = 0.0001}$	160.0v	82v	26.9v	3.0v	0.1v

Perhaps an example will make this result more intuitive.[6] Consider two cases:

Case 1: For each day this weekend, you have a 20% daily chance of dying. However, if you survive the weekend you are likely to live a long and happy life.

Case 2: For each day this year, you have a 20% daily chance of dying. However, if you survive the next year, you are likely to live a long and happy life.

In Case 1, it makes sense to invest significant resources in reducing the chance that you will die over the weekend. But in Case 2, you are quite unlikely to survive the next year no matter what you do. In this case, it makes more sense to invest in making your life during this year as happy as possible. A long time of perils means that humanity's future is closer to Case 2 than to Case 1, so that the value of existential risk mitigation is unlikely to be astronomical.

[6] Thanks to an anonymous referee for suggesting this example.

Where does this discussion leave the pessimistic case for the astronomical value thesis? It is time to take stock.

3.5 Taking stock

In this section, we considered four ways of modifying the Simple Model to support the astronomical value thesis. We saw that a distinction between absolute and relative risk can only harm the pessimist. We also saw that neither optimistic growth assumptions, nor even the assumption that actors can affect risk across all centuries at once will be sufficient to ground the astronomical value thesis. And we saw that the most likely culprit for these failures is pessimism itself.

We saw that the best way to reconcile pessimism with the astronomical value thesis involves a strong empirical assumption. This is the time of perils hypothesis on which existential risk will be high during the coming centuries, but then drop to a much lower level of post-peril risk if we survive this perilous period. We saw that the time of perils hypothesis could well bear out the astronomical value of existential risk reduction, provided two conditions hold: the perilous period is short, and the level of post-peril risk is very low. But should we believe this version of the time of perils hypothesis?

To see the gap between pessimism and the time of perils hypothesis, consider the pessimist's reasons for thinking that existential risk is currently high. Pessimists think that existential risk is high because we have developed new technologies with unprecedented destructive potential. However, future technology is likely to far outstrip our own, so this same argument might be taken to suggest that future risk will be higher, not lower than current risk. If the pessimist is to resist this conclusion, she needs to argue that humanity will soon learn to effectively manage the risks posed by new technologies. In the rest of this paper, I consider three arguments that have been advanced for that conclusion and argue that they are unlikely to ground a time of perils hypothesis of the needed form.[7]

4. Wisdom

Sagan took the problem to be that humanity's technological capabilities are growing far more quickly than our wisdom. Until we gain the wisdom to handle new technologies, Sagan held, we will remain at peril. But once we grow in wisdom, we may become relatively safe.

This line has been taken up by other pessimists. Here is Ord, quoting Sagan:

> The problem is not so much an excess of technology as a lack of wisdom. Carl Sagan put this especially well: "Many of the dangers we face indeed arise from science and technology — but, more fundamentally because we have become powerful without becoming commensurately wise." (Ord 2020, p. 45).

Sagan put a sharper edge on the point: "If we continue to accumulate only power and not wisdom, we will surely destroy ourselves" (Sagan 1997, p. 185).

The trouble with this argument is that it is thin on details. Neither Sagan nor Ord tells us much about what it means to become wise; why we should expect dramatic future increases in wisdom; and how increased wisdom could lead to a short perilous period followed by a dramatic reduction in post-peril risk. There are interesting ways of

[7] One argument which I will not address is that the development of artificial intelligence may bring an end to the time of perils, for example by putting human civilization under the control of a single entity capable of managing existential risks (Bostrom 2014). The response to this argument turns on a number of conceptual and empirical questions surrounding artificial intelligence that are difficult to address in the space of a paper.

precisifying the argument, but none of them will ground a time of perils hypothesis of the right form.

One thing we might do is to point towards promising current trends in reasoning and related areas. In this vein, Nick Bostrom argues that:

> An optimist could expect that the 'sanity level' of humanity will rise over the course of this century — that prejudices will (on balance) recede, that insights will accumulate, and that people will become more accustomed to thinking about abstract future probabilities and global risks. With luck, we could see a general uplift of epistemic standards in both individual and collective cognition. (Bostrom 2014, p. 284).

It may not be unreasonable to hope for the uplift in epistemic standards that Bostrom describes. But the problem is that these and similar trends come nowhere close to grounding the manyfold reduction in post-peril risk that the pessimist needs. It is not so implausible to think that a reduction in prejudice or a rise in future-oriented thinking might lead humanity to take existential risks more seriously. But these are moderate and familiar trends, and on their own they are highly unlikely to be strong enough to take us out of the time of perils. Indeed, it is perhaps for this reason that Bostrom hedges his appeal to increased sanity by attributing this thought to an optimist and does not saddle even the optimist with the claim that increased sanity alone will be powerful enough to take us out of the time of perils.

Ord (2020) strengthens Bostrom's argument by appealing to civilizational virtues. Ord argues that we can treat humanity as a collective agent currently in its infancy.

Humanity will grow in wisdom and reach adulthood by acquiring civilizational virtues such as prudence, patience, self-discipline, compassion, stewardship, gratitude, fairness, unity and solidarity. As humanity grows in virtue and hence in wisdom, humanity will act to substantially reduce existential risk, bringing an end to the time of perils. This view strengthens Bostrom's argument by divorcing the concept of civilizational virtue from the virtues of individual humans. Because collective agents can have properties that their members lack, Ord holds, we may well hope that humanity as a whole will become substantially more patient or compassionate in the coming centuries, even if we doubt that the average human will grow in patience or compassion during this time.

At this point, the most helpful response would be to ask Ord for more details. We are not told much about why we should expect humanity to grow in virtue or how this growth could lead to a quick and substantial drop in existential risk. Without these details, it is hard to place much stock in the appeal to civilizational virtue. But we may get some handle on the prospects for Ord's argument by thinking through some particular civilizational virtues.

Consider unity. Humanity becomes more unified as we build forms of international cooperation such as the United Nations, or international trade and climate agreements. Becoming unified ensures that humanity acts with a view to the interests of humanity as a whole, instead of each nation pursuing its own interest. This would increase pressure to address existential risks, since humanity would be concerned with the security of all humans and their descendants, instead of the security of a single nation and its descendants. But increased unity could only do so much to drive down existential risk. At the time of writing, many nations boast at least 5% of the world's population within their own borders and at least two contain over 15% of the world's population. Unifying these nations into a single actor would increase their constituencies by a factor of no more than twenty, and hence in the best case it could not lead to more than a twentyfold increase in the importance of existential risk reduction. While that is nothing to sneeze at, it remains orders of magnitude lower than what the pessimist needs.

Next, consider patience. Humanity becomes more patient by adopting systems of government which better represent the interests of future people. Many political systems give inadequate weight to future generations, for example by instituting short election cycles which force politicians to deliver immediate results, or by giving no formal voice to unborn generations (Thompson 2010). These problems can, and have been partially addressed by mechanisms such as citizens' assemblies elected to represent future generations, or government commissioners tasked with protecting future generations (John and MacAskill 2021).

There is no doubt that institutional changes can increase the patience of political systems. For example, many of these changes have led to increased emphasis on mitigating climate risks to future generations. But the pessimist needs a mechanism by which a largely impatient group of humans could together become patient enough to make great sacrifices directed at reducing existential risks, based largely on the threat that those risks pose to far-future generations. It is hard to see how the features of current, or feasible near-term political changes could increase patience on such a scale. Indeed, one might reasonably expect constituents to reject any system of government that acted with substantially more patience than the average voter.

Now it could well be that there is a plausible story about how humanity might acquire some particular virtue that is strong enough to end the time of perils. Or perhaps the combination of many different virtues will be enough to tip the scales. But we have not seen a detailed argument for either of these conclusions, and we saw above that making the argument out is no easy task. So we cannot yet ground the time of perils hypothesis in the hope that humanity will increase in wisdom. In the next section, I consider an economic argument for the time of perils hypothesis.

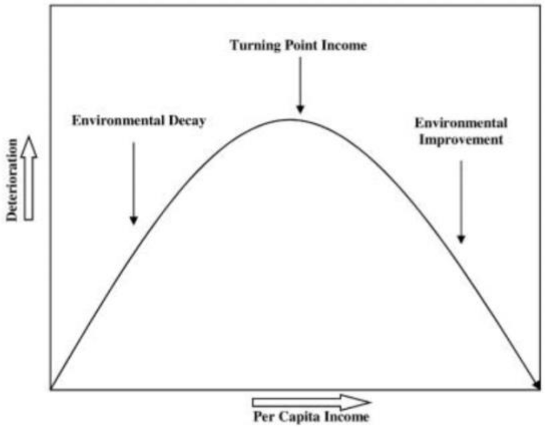
5. An existential risk Kuznets curve?

Consider the risk of climate catastrophe. Climate risk increases with growth in consumption, which emits fossil fuels. Climate risk decreases with spending on climate safety, such as reforestation and other forms of carbon capture.

Economists have noted that two dynamics exert pressure towards reduced climate risk in sufficiently wealthy societies. First, the marginal utility of additional consumption decreases, reducing the benefit to fossil fuel emissions. Second, as society becomes wealthier we have more to lose by destroying our climate. These dynamics exert pressure towards an increase in safety spending relative to consumption.

Some economists have hypothesized that this dynamic is sufficient to generate an *environmental Kuznets curve* (Figure 1): an inverse U-shaped relationship between per-capita income and environmental degradation (Dasgupta et al. 2002; Grossman and Krueger 1995; Stokey 1998). Societies initially become wealthy by emitting fossil fuels and otherwise degrading their environments. But past a high threshold of wealth, rational societies should be expected to improve the environment more quickly than they destroy it, due to the diminishing marginal utility of consumption and the increasing importance of climate safety.

Figure 1: The environmental Kuznets curve. Reprinted from (Yandle et al. 2002).



Now it is widely conceded that this dynamic will not be fast enough to stop the world from causing irresponsible levels of environmental harm. But it may well be enough to prevent the most catastrophic warming scenarios, where 10-20°C warming may lead to human extinction or permanent curtailment of human potential.[8]

Leopold Aschenbrenner (2020) argues that the same dynamic repeats for other existential risks. Aschenbrenner's argument draws on a Solow-style growth model (Solow 1956) extending Jones (2016). In this model, society is divided into separate consumption and safety sectors. At time t, the consumption sector produces consumption outputs C_t as a function of the current level of consumption technology A_t and the labor force producing consumption goods L_{ct}.

[8] These drastic scenarios might require burning the entire stock of fossil fuels on Earth (Tokarska et al. 2016). Even then, (Ord 2020) notes, these scenarios may well stop short of existential catastrophe.

$$C_t = A_t^{\alpha} L_{ct}.$$

Here $\alpha > 0$ is a constant determining the influence of technology on production.

Similarly, the safety sector produces safety outputs H_t as a function of safety technology B_t and the labor force producing safety outputs L_{ht}.

$$H_t = B_t^{\alpha} L_{ht}.$$

As in the environmental case, Aschenbrenner takes existential risk δ_t to increase with consumption outputs and decrease with safety outputs. In particular, he assumes:

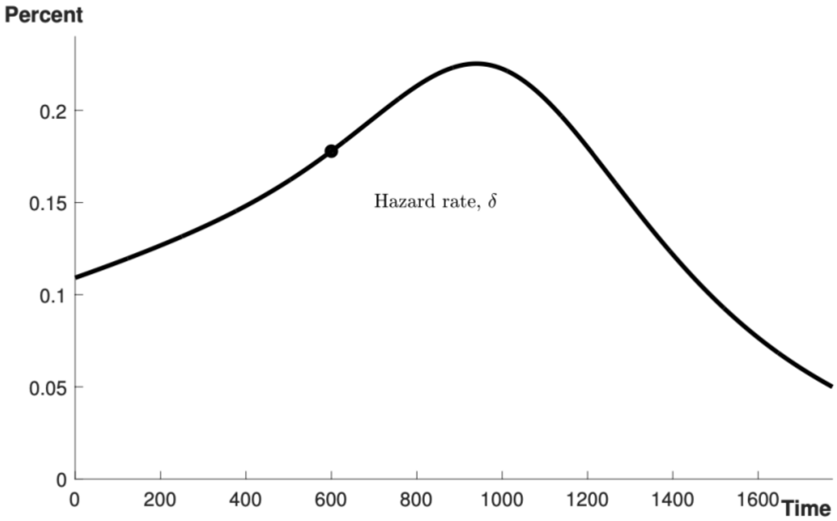
$$\delta_t = \overline{\delta} C_t^{\epsilon} H_t^{-\beta}.$$

for constants $\overline{\delta}, \epsilon, \beta$.

Aschenbrenner proves that under a variety of conditions, optimal resource allocation should lead society to invest quickly enough in safety over consumption to drive existential risk towards zero. Roughly put, if the marginal utility of consumption falls quickly enough and if consumption outputs do not increase existential risk much more quickly than safety outputs decrease risk, then optimal resource allocation leads to a nontrivial probability of humanity surviving for billions of years.

Aschenbrenner shows that under a range of assumptions, his model grounds an *existential risk Kuznets curve*: a U-shaped relationship between time and existential risk (Figure 2). Although existential risk remains high today and may increase for several centuries, eventually the diminishing marginal utility of consumption and the increasing importance of safety should chase risk exponentially towards zero. Until that happens, humanity remains in a time of perils, but afterwards, we should expect low levels of post-peril risk to continue indefinitely into the future.

Figure 2: The existential risk Kuznets curve. Reprinted from (Aschenbrenner 2020).



I think this is among the best arguments for the time of perils hypothesis. At the same time, I have two doubts about this form of the argument. First, the Aschenbrenner model treats consumption as the driver of existential risk. But most pessimists do not think that consumption is even the primary determinant of existential risk. In the special case of climate risk, consumption does indeed drive risk by emitting fossil fuels and causing other forms of environmental degradation. But pessimists think that the lion's share of existential risk comes from risks such as rogue artificial intelligence and sophisticated bioterrorism. These risks are not caused primarily by consumption, but rather by technological growth. Risks from superintelligence grow with advances in technologies such as machine learning, and bioterrorism risks grow with advances in our capacity to synthesize, analyze and distribute biological materials. So a reduction in existential risk may be largely achieved through slowing growth of technology rather by slowing consumption.

We could revise (3) to let technologies A and B replace consumption outputs C as the main drivers of existential risk. But technology occupies a very different role from consumption outputs in the Aschenbrenner model. One difference is that technology is an input rather than an output to production in (1) and (2). In general we have no reason to expect symmetrical results to govern inputs and outputs in mathematical models, hence we have no good reason to expect results proved for consumption outputs to generalize to technology.

Another difficulty is that technology governs both the safety and consumption sectors, whereas consumption outputs have no direct bearing on safety outputs. This is important, because the proofs of Aschenbrenner's main results rely on the idea that societies can sharply curtail existential risk by devoting increasing amounts of labor and scientific research to the safety sector. But increased labor alone is often insufficient to guarantee safety given current technologies, and new safety technologies may themselves carry risk. When this is the case, it is not so clear that we can significantly reduce risk by shifting labor and research from the consumption sector to the safety sector.

For example, one risk discussed by pessimists is the risk of asteroid impacts (Bostrom 2013; Ord 2020). There is mounting evidence that an asteroid impact during the Cretaceous period wiped out every land-dwelling mammal weighing more than five kilograms (Alvarez et al. 1980; Schulte et al. 2010), and a similar impact could well extinguish humanity. It is widely accepted that increased labor, given current technology, cannot eliminate risks from asteroid impacts. While there are some things we can do to promote safety given current technology, such as stockpiling food, full safety would require the capacity to deflect large incoming asteroids. Developing this capacity would require research into deflection technologies. But in fact, leading pessimists think that researching asteroid deflection technologies would be a bad idea (Ord 2020). Deflection technologies are likely to be used for mining and military applications, and those applications carry a higher risk of deflecting asteroids towards Earth than away from Earth. Here we have a case where existential risk cannot be substantially reduced by reallocating labor to the safety sector and in which safety research may increase rather than decrease existential risk. Cases such as this one put pressure on the idea that we can produce a manyfold reduction in existential risk by reallocating labor and technological research from the consumption to safety sectors.

So far, we have discussed a series of technical worries for Aschenbrenner's result. Aschenbrenner's model takes consumption outputs, rather than technology to be the primary driver of existential risk. Changing this assumption casts doubt on the Aschenbrenner result, since technology is an input rather than an output of production and because existential risks brought about by safety technologies may be significant.

A different worry is that the Aschenbrenner result holds when resources are allocated optimally. As Aschenbrenner notes, this may not be the case. For one thing, safety is a global public good and theory predicts that global public goods will be sharply undersupplied (Ord 2020). Even the largest countries bear only a fraction of global risk burdens, and each nation would prefer to leave existential risk reduction to others. Moreover, much of the disvalue of existential risks comes in their impact on the distant future, and there are good reasons to expect that far-future value will be underpromoted. Indeed, pessimists think that existential risk mitigation has been radically underfunded to date. Aschenbrenner's model does address one reason why future value may be neglected, namely a positive rate of pure time preference. But it does not address the many other motivational and institutional obstacles, such as cognitive biases and short-term election cycles, which are often held up as obstacles to longtermist political decisionmaking (John and MacAskill 2021). For these reasons, we might worry that even if an optimal resource allocation would bring an end to the time of perils, human societies may suboptimally allocate resources away from existential risk mitigation at the expense of continued peril.

In this section, I considered an argument due to Leopold Aschenbrenner for the time of perils hypothesis based on the idea of an existential risk Kuznets curve. I argued that the Aschenbrenner model assumes, unlike leading pessimists, that consumption rather than technological growth drives existential risk, and that once this assumption is removed the argument runs into trouble on two fronts. I also raised worries for Aschenbrenner's assumption of optimal resource allocation. Together, I think these arguments cast some doubt on the idea that the time of perils hypothesis can be defended by positing an existential risk Kuznets curve. In the next section, I consider one final argument for the time of perils hypothesis.

6. Settling the stars

It is sometimes proposed that the time of perils will end as human civilization expands throughout the stars.[9] So long as humanity remains tied to a single planet, we can be wiped out by a single catastrophe. But if humanity settles many different planets, each with its own land mass, values and system of government, our geographic, institutional and

[9] Not all pessimists are convinced (Ord 2020). And more generally some have argued that space settlement *increases* existential risk, for example by contributing to military conflict (Deudney 2020; Torres 2018).

cultural diversity may provide good insurance against the spread of catastrophes from one planet to another. Then it might take an unlikely sequence of independent calamities to present a permanent challenge to the survival or development of human civilization as a whole.

This thought has been defended at length by the astronomer Martin Ćirković (2019), who argues that space colonization is the only viable prospect for long-term human survival. It was also voiced by Sagan, who concluded immediately after introducing the concept of the time of perils that "every surviving civilization is obliged to become spacefaring — not because of exploratory or romantic zeal, but for the most practical reason imaginable: staying alive" (Sagan 1997, p. 173). And the same thought has been cited by Elon Musk as one of his primary reasons for pursuing Mars colonization (Musk, 2017). Could the prospects for space settlement ground a time of perils hypothesis of the needed form?

This is unlikely. To see the problem, distinguish two types of existential risks: *anthropogenic* risks posed by human activity such as greenhouse gas emissions and bioterrorism, and *natural risks* posed by the environment, such as asteroid impacts, supervolcanoes or naturally occurring diseases. Advocates of space settlement have rightly noted that settling the stars could greatly reduce the risk of existential catastrophe from natural causes (Gottlieb 2019; Schwartz 2011). It is exceedingly unlikely for events such as asteroid impacts or supervolcanoes to strike two planets in quick succession. But pessimists think that the most pressing existential risks are anthropogenic risks, rather than natural risks. By way of illustration, Ord (2020) estimates natural risk in the next century at one in ten thousand, but overall existential risk this century at one in six. So a reduction in natural risk is cold comfort to the pessimist, and nothing like the sharp drop in post-peril risk that she needs.

Could settling the stars bring quick relief for anthropogenic risks? Perhaps space settlement would help with some anthropogenic risks, such as the risks posed by climate change. But these risks are not the major drivers of existential risk pessimism. Ord (2020) estimates existential risk from climate change in the next century at one in a thousand. Many pessimists, including Ord, think that a large fraction of anthropogenic risk is driven by risks from bioterrorism and the use of artificial intelligence systems whose goals are misaligned with our own. Could space settlement mitigate such risks?

Perhaps there is a sense in which our very distant descendants may be protected from such risks after they have settled many star systems. For example, Ćirković (2019) notes that if the laws of physics prohibit faster-than-light travel, then a human civilization spread over many light years may have years to prepare against the spread of catastrophes between systems. But this will not come about for many millennia, so it is of little help to the pessimist who needs to argue that the time of perils will be short.

In the short-term, it is hard to see how feasible levels of space settlement could protect against risks such as bioterrorism and misaligned artificial intelligence. Are we to imagine that a superintelligent machine could come to control all life on Earth, but find itself stymied by a few stubborn Martian colonists? That a dastardly group of scientists designs and unleashes a pandemic which kills every human living on Earth, but cannot manage to transport the pathogen to other planets within our solar system? Perhaps there is some probability of such scenarios, but they hardly ground the manyfold drop in post-peril risk that the pessimist needs.

In this section, we have seen that an appeal to space settlement is unlikely to ground the time of perils hypothesis. While space settlement may do much to mitigate natural risks, these risks play only a small part in existential risk pessimism. By contrast, it is hard to see how space settlement could quickly and effectively mitigate the anthropogenic risks underlying pessimistic estimates of current existential risk.

7. Philosophical implications

This paper explored the relationship between existential risk pessimism and the astronomical value thesis. Section 2 explored a Simple Model of existential risk on which the value of existential risk mitigation may be relatively modest, and is unaffected by existential risk pessimism. Section 3 considered four extensions of the Simple Model. These models suggested that existential risk pessimism may lower rather than raise the value of existential risk mitigation. The first three models also failed to support the astronomical value thesis.

The best way out for the pessimist, I suggested, is to invoke the time of perils hypothesis on which existential risk is high now, but will shortly fall to a low level. I argued that the time of perils hypothesis could well ground the astronomical value thesis, so long as the perilous period is short and the post-peril risk is low. However, Sections 4-6 considered three arguments for the time of perils hypothesis and found these arguments to be inconclusive. This discussion has at least four important philosophical implications which are worth exploring.

7.1 The demandingness of consequentialism

A common complaint against consequentialism is that consequentialism is too demanding.[10] Traditionally, the problem is expressed in the following way: consequentialism demands a great degree of personal sacrifice by present individuals in order to benefit other present individuals. For example, we might be obligated to donate a good deal of our savings to benefit the poor (Pummer 2016; Singer 1972). This strikes some authors as implausible, because it makes actions which seem permissible but supererogatory into obligatory actions (Scheffler 1994).

Some recent authors have thought that consequentialism is even more demanding than this (Mogensen 2021). Given the scale of humanity's future, consequentialism seems to require present people to use a good deal of our wealth to benefit future people. If existential risk is high, we should use that wealth to mitigate existential risk, and if existential risk is low, we should invest our wealth to build institutions and economic resources for the future (Trammell 2021). This may strike readers as even less plausible, because it involves neglecting important contemporary problems such as racial injustice or global poverty. Some readers may think that funneling resources away from these contemporary problems is neither obligatory nor even supererogatory, but instead flatly impermissible.

One implication of this paper is that existential risk pessimism tends to favor the original statement of the demandingness problem. If existential risk is indeed high, then it is relatively less important to mitigate existential risk, and it will be also relatively less important to save resources for the future, since future gains are less likely to be realized.

[10] For discussion see Kagan (1984), Mulgan (2001) and Sobel (2007).

That means it may indeed be better for consequentialists to direct their resources towards present people. This should be welcome news for the consequentialist, since it raises the possibility of avoiding a strengthened form of the demandingness objection to consequentialism.

7.2 Longtermism

An increasing number of philosophers identify as longtermists (Greaves and MacAskill 2021; [removed]; Ord 2020; Tarsney 2022). As an axiological matter, longtermism holds roughly that the best thing we can do today is to act to benefit our far-future descendants. A bit more precisely, Hilary Greaves and Will MacAskill (2021) define axiological longtermism as follows:

> **Axiological strong longtermism (ASL):** In the most important decision situations facing agents today,
>
> (i) Every option that is near-best overall is near-best for the far future.
>
> (ii) Every option that is near-best overall delivers much larger benefits in the far future than in the near future.

In addition to its philosophical importance, longtermism has had tremendous public impact through popular publications (Ord 2020; MacAskill 2022). It has also exerted a substantial influence on charitable giving, drawing money away from areas such as global health and poverty reduction and towards causes such as existential risk mitigation.

Many arguments for ASL cite existential risk mitigation as among the best motivations for ASL (Greaves and MacAskill 2021; Bostrom 2013). Many longtermists are also pessimists about current levels of existential risk. For example, we saw that Toby Ord posits a 1/6 chance of existential catastrophe by 2100. One of the lessons of this paper is that existential risk pessimism tends to significantly lower the value of existential risk mitigation. If we reject the time of perils hypothesis but remain pessimistic about levels of existential risk, it may well turn out that the value of existential risk mitigation is insufficient to ground ASL in many contemporary decision problems. This would rule out a popular argument for ASL.

More generally, [removed] has suggested that longtermists face an *optimism/pessimism dilemma*. If longtermists are optimistic about the value of the future, then it becomes harder to identify tractable ways to improve the future. But if longtermists are pessimistic about the value of the future, then it becomes less important to preserve the future. The argument in this paper provides some support for the pessimism horn of [removed]'s dilemma. Existential risk pessimism is one of the most common ways for longtermists to be pessimistic about the value of the future. We have seen that existential risk pessimism strongly reduces the value of existential risk mitigation. Future work might explore the impact of existential risk pessimism on other efforts beyond existential risk mitigation to improve the long-term future. Plausibly, the models in this paper should exert significant downward pressure on the values of those actions due to uncertainty about the length of the future that is being improved by our actions. In this way, it may be productive to interpret this paper as an avenue of support for the pessimism horn of [removed]'s optimism/pessimism dilemma.

7.3 Discounting

A common way to prevent long-term value from gaining undue importance in axiological discussions is to apply a discount rate whereby future value is discounted relative to present value. Discount rates of several percent per year are standardly applied to value quantities such as the harms caused by climate change (Dasgupta 2008) and the benefits of policy interventions (Greaves 2017).

Beyond standard economic sources of diminishing value, such as quality depreciation or monetary inflation, there are two common ways to justify discounting future value. First, we may discount future value due to uncertainty: future gains are devalued because we are unsure whether we will actually realize them.[11] Second, we may discount future value due to pure time preference, the bare axiological view that present benefits matter more than future benefits because of their temporal location.

Although pure time preference is popular in some disciplines, many philosophers find pure time preference difficult to justify (Broome 1994; Gardiner 2004). For many philosophers, it is hard to see how the mere temporal location of benefits could make those benefits matter more or less. However, appeals to pure time preference (Lloyd 2021) or close analogs thereof (Mogensen 2022) have recently surfaced as strategies for resisting axiological longtermism. Here the thought may be that uncertainty-based discounting alone would not be enough to push far-future benefits to a level comparable to short-term benefits.

The discussion in this paper suggests a route by which some opponents of longtermism could make do with uncertainty-based discounting without appeal to pure time preference. In the models of Sections 2-3, high rates of existential risk generate

[11] See Gabaix and Laibson (2022) for a good explanation of the relationship between uncertainty and discounting.

substantial uncertainty about whether future value will be realized, because we are uncertain about whether there will be humans around to realize this value. Although these models make no appeal to pure time preference, we saw that on their own, these models substantially discount future value when combined with pessimistic views about existential risk. This discussion shows how existential risk pessimism could be used to replace pure time preference in discounting far-future value, even under relatively optimistic assumptions about the amount of value that future centuries could hold.

7.4 The moral importance of existential risk mitigation

Where does this discussion leave the case for existential risk mitigation? To some extent, this depends on readers' views about the cost-effectiveness of existing opportunities to mitigate existential risks. Perhaps some efforts to mitigate existential risks are so cost-effective that they can be justified only by their benefits for agents alive today ([removed]). The arguments of this paper will do little to challenge the value of such interventions. More generally, we might sympathize with Ord (2020), who bemoans the fact that the world spends more on ice cream than on existential risk mitigation. Existential risk mitigation could turn out to be, on the current margin, quite valuable. But in general, the models of this paper suggest that existential risk mitigation may not be as important as many pessimists have taken it to be, and crucially that pessimism is a hindrance rather than a support to the case for existential risk mitigation. The case for existential risk mitigation is strongest under more optimistic assumptions about existential risk.

Appendix A: Proofs of results

The simple model

(Simple Model) $V[W] = v\sum_{i=1}^{\infty}(1-r)^i$.

Note that $V[W]$ is a truncated geometric series so that:

$$V[W] = v\left(\frac{1}{1-(1-r)} - 1\right) = v\frac{1-r}{r}.$$

Let X be an intervention reducing risk in this century to $(1-f)r$, and let W_X be the result of performing X. Then

$$\begin{aligned} V[W_X] &= v(1-(1-f)r)\sum_{i=1}^{\infty}(1-r)^{i-1} \\ &= v(1-(1-f)r)\left(\frac{1}{1-(1-r)}\right) \\ &= v\frac{(1-(1-f)r)}{r}. \end{aligned}$$

And hence:

$$\begin{aligned} V[X] &= V[W_X] - V[W] \\ &= v\frac{(1-(1-f)r)}{r} - v\frac{1-r}{r} \\ &= fv. \end{aligned}$$

Absolute risk reduction

Let X_{ABS} be an intervention reducing risk in this century to $r-f$ for $f \leq r$. Then:

$$\begin{aligned} V[W_X] &= v(1-(1-f)r)\sum_{i=1}^{\infty}(1-r)^{i-1} \\ &= v\frac{1-(r-f)}{r}. \end{aligned}$$

So that:

$$\begin{aligned} V[X] &= V\left[W_{X_{ABS}}\right] - V[W] \\ &= v[\frac{1-(r-f)}{r} - \frac{1-r}{r}] \\ &= fv/r. \end{aligned}$$

Linear growth

(Linear Growth) $V[W] = v \sum_{i=1}^{\infty} i\,(1-r)^{i}$.

Note that $V[W]$ is a polylogarithm with order -1. Recalling that

$$\sum_{i=1}^{\infty} \frac{z^{i}}{i^{-1}} = \frac{z}{(1-z)^{2}}.$$

we have:

$$V[W] = v\frac{1-r}{[1-(1-r)]^{2}} = v(1-r)/r^{2}.$$

If X produces a relative reduction of risk by f then:

$$\begin{aligned} V[W_X] &= v(1-(1-f)r)\sum_{i=1}^{\infty} i\,(1-r)^{i-1} \\ &= v\frac{1-(1-f)r}{1-r}\sum_{i=1}^{\infty} i\,(1-r)^{i} \\ &= v\left(\frac{1-(1-f)r}{1-r}\right)\left(\frac{1-r}{r^{2}}\right) \\ &= v\frac{1-(1-f)r}{r^{2}}. \end{aligned}$$

So that:

$$\begin{aligned} V[X] &= V[W_X] - V[W] \\ &= v\frac{1-(1-f)r}{r^2} - v\frac{1-r}{r^2} \\ &= fv/r. \end{aligned}$$

Quadratic growth

(Quadratic Growth) $V[W] = v\sum_{i=1}^{\infty} i^2\,(1-r)^i$.

Note that $V[W]$ is a polylogarithm with order -2. Recalling that

$$\sum_{i=1}^{\infty}\frac{z^i}{i^{-2}} = \frac{z(1+z)}{(1-z)^3}$$

we have

$$V[W] = v\frac{(1-r)\big(1+(1-r)\big)}{\big(1-(1-r)\big)^3} = v\frac{(1-r)(2-r)}{r^3}.$$

With X as before we have:

$$\begin{aligned} V[W_X] &= v(1-(1-f)r)\sum_{i=1}^{\infty} i^2\,(1-r)^{i-1} \\ &= v\frac{1-(1-f)r}{1-r}\sum_{i=1}^{\infty} i^2\,(1-r)^i \\ &= v\left(\frac{1-(1-f)r}{1-r}\right)\left(\frac{(1-r)(2-r)}{r^3}\right) \\ &= v\frac{[1-(1-f)r](2-r)}{r^3}. \end{aligned}$$

This gives:

$$\begin{aligned} V[X] &= V[W_X] - V[W] \\ &= v\frac{[1-(1-f)r](2-r)}{r^3} - v\frac{(1-r)(2-r)}{r^3} \\ &= \left(\frac{v(2-r)}{r^3}\right)[1-(1-f)r-(1-r)] \\ &= \left(\frac{v(2-r)}{r^3}\right)(fr) \\ &= fv(2-r)/r^2. \end{aligned}$$

Global risk reduction

If X produces a global (relative) reduction in risk by f, then

$$V[W_X] = v\sum_{i=1}^{\infty}(1-(1-f)r)^i = v\frac{1-(1-f)r}{(1-f)r}.$$

so that

$$\begin{aligned} V[X] &= V[W_X] - V[W] \\ &= v\frac{1-(1-f)r}{(1-f)r} - v\frac{1-r}{r} \\ &= \left(\frac{v}{r}\right)\left(\frac{1-(1-f)r-(1-f)(1-r)}{1-f}\right) \\ &= \frac{v}{r}\frac{f}{1-f}. \end{aligned}$$

The time of perils

(Time of Perils) $V[W] = \sum_{i=1}^{N} v\,(1-r)^i + (1-r)^N \sum_{i=1}^{\infty} v\,(1-r_l)^i$.

Note that:

$$\begin{aligned} V[W] &= v\left[\frac{1-(1-r)^{N+1}}{1-(1-r)} - 1\right] + (1-r)^N v\frac{1-r_l}{r_l} \\ &= v\frac{1-r}{r}[1-(1-r)^N] + (1-r)^N v\frac{1-r_l}{r_l}. \end{aligned}$$

If X leads to a relative reduction of risk by f in the next century, then:

$$V[W_X] = (1-(1-f)r)[v\sum_{i=1}^{\mathrm{N}}(1-r)^{i-1} + (1-r)^{N-1}v\sum_{i=1}^{\infty}(1-r_l)^{i}]$$

$$= \; v(1-(1-f)r)[\frac{1-(1-r)^{N}}{r} + (1-r)^{N-1}V_{\mathrm{SAFE}}].$$

Subtracting term-wise gives:

$$V[X] = V[W_X] - V[W]$$

$$= \frac{v[1-(1-r)^{N}]}{r}[1-(1-f)r-(1-r)] +$$

$$V_{\mathrm{SAFE}}[(1-(1-f)r)(1-r)^{N-1}-(1-r)^{N}]$$

$$= \; fv[1-(1-r)^{N}] + (1-r)^{N-1}V_{\mathrm{SAFE}}[1-(1-f)r-(1-r)]$$

$$= \; fv[1-(1-r)^{N}] + fr(1-r)^{N-1}V_{\mathrm{SAFE}}.$$

Appendix B: Rapid value growth

Some readers may be interested in examining whether the results of Section 3.2 generalize to more extreme forms of value growth. The models in Section 3.2 are already more generous than their immediate predecessors: both Ord and Adamcewski consider only linear growth, whereas Section 3.2 extends their results to the case of quadratic growth. However, in this appendix, I consider extensions to cubic and exponential growth.

Cubic growth

Some readers may think that the value of future centuries grows cubically in the number of elapsed centuries. For example, Christian Tarsney (2022) considers a scenario in which humanity begins traveling the stars at a fixed speed in all directions at once, colonizing all habitable planets in our path. So long as this period of aggressive expansion continues, we will see cubic growth in the area of settled space, which could be argued to ground cubic growth in the value of future centuries.

I would like to urge caution in modeling indefinite cubic value growth. One reason for caution is that our best demographic models project that the future human population will shrink rather than grow, and that the age of Malthusian expansion in which rapid resource acquisition allows us to quickly grow the human population is behind us (Eden and Alexandrie forthcoming, Geruso and Spears forthcoming). Another caution is that even a spacefaring population would have to be quite aggressively focused on expansion to find itself setting sail in every direction at once for the purpose of colonization, maintaining this pattern of expansion over many centuries. Finally, as we saw in this paper, the timing of expansion matters: if aggressive interstellar expansion is meant to ground cubic value growth, then that period of expansion had better begin soon, or it will not be much help to the value of existential risk mitigation.

At the same time, cubic growth leads to a broadly similar pattern as the models of Section 3.2.

(Cubic Growth) $V[W] = v \sum_{i=1}^{\infty} i^3 \, (1-r)^i$.

Note that $V[W]$ is a polylogarithm with order -3. Recalling that:

$$\sum_{i=1}^{\infty} \frac{z^i}{i^{-3}} = \frac{z(1+4z+z^2)}{(1-z)^4}$$

we have:

$$V[W] = \frac{(r^2-6r+6)(1-r)}{r^4}.$$

If X leads to a relative reduction of risk by f in the next century, then:

$$\begin{aligned} V[W_X] &= (1-(1-f)r)(v + v\sum_{i=1}^{\infty}(i+1)^3\,(1-r)^i) \\ &= (1-(1-f)r)(v + \left(\frac{v}{1-r}\right)\sum_{i=1}^{\infty}(i+1)^3\,(1-r)^{i+1}) \\ &= (1-(1-f)r)(v + \frac{1}{(1-r)}(V[W] - v(1-r)) \\ &= \frac{1-(1-f)r}{1-\mathrm{r}}V[W] \end{aligned}$$

So that:

$$\begin{aligned} V[X] &= V[W_X] - V[W] \\ &= \mathrm{V[W]}\left(\frac{(1-(1-\mathrm{f})\mathrm{r})}{1-\mathrm{r}} - 1\right) \\ &= \mathrm{V[W]}\frac{fr}{1-r} \\ &= \frac{f(r^2\ - 6r + 6)}{r^3}. \end{aligned}$$

As before, we see that pessimism decreases rather than increases the value of existential risk mitigation. V[X] is bounded above by $7/r^3$, so that to a crude approximation V[X] decays cubically in the level r of starting risk. And as before, we cannot get pessimism and the astronomical value thesis to be true together. Table 3 illustrates these lessons by extending Table 1 to cover the case of cubic value growth.

Table 3: Value of 10% relative risk reduction across growth models and risk levels

	r = 0.2	**r = 0.02**	**r = 0.002**	**r = 0.0002**
Linear growth	0.5v	5v	50v	500v
Quadratic growth	4.5v	495v	49,950v	$5{*}10^6$v
Cubic growth	60.5v	73,505v	$7.5{*}10^7$v	$7.5{*}10^{10}$v

These values will be significantly reduced if the onset of cubic growth is delayed. See Tarsney (2022) for a complementary model with delayed onset of cubic growth.

Exponential growth

Some readers may wonder whether the results of Section 3.2 generalize to the case of indefinite *exponential* growth in the value of a century. The most important thing to say about exponential models of value growth is that they are implausible. Even when we look at growth in economic outputs such as GDP, a few recent centuries of exponential growth present as an island among millions of years of previous subexponential growth. While many economists think that exponential GDP growth may continue for some time yet, no economists project that exponential growth will continue indefinitely, because exponential functions grow far too quickly in the long run. For example, projecting forward an estimated 2.1% annual growth in per-capita GDP during this century (Christensen et al. 2018) into the indefinite future, starting from today's per-capita world GDP of $12,000, would yield a per-capita GDP above $$10^{100}$ within 110 centuries. That is a very large number.

Another reason to be skeptical of indefinite exponential growth is that exponential economic growth need not translate into exponential value growth. From the fact that Elon Musk is over a million times richer than I am, it does not follow that Musk's well-being is a million times higher than mine. The most plausible way to recover exponential value growth from exponential economic growth would be to posit a tight Malthusian link between GDP and population, but as we saw above, most economists think that this Malthusian regime is gone and may never return (Eden and Alexandrie forthcoming, Geruso and Spears forthcoming).

Readers interested in exponential growth models may want to consult Aschenbrenner (2020). Alternatively, here is an exponential growth model in the spirit of this paper.

Exponential growth values the world at:

$$V[W] = v \sum_{i=1}^{\infty} (a(1-r))^i$$

where a is the rate of per-century value growth. This diverges for a ≥ 1/(1-r) and otherwise converges to:

$$V[W] = \frac{va(1-r)}{1-a(1-r)}.$$

In the convergent case, if X leads to a relative reduction of risk by f in the next century, then:

$$V[W_X] = (1-(1-f)r)[av + av\sum_{i=1}^{\infty}\left(a(1-r)\right)^i] = a(1-(1-f)r)(v + V[w]).$$

And hence:

$$V[\mathrm{X}] = V[W_X] - V[\mathrm{W}] = \ V[w](a(1-(1-f)r)-1) + a(1-(1-f)r)v.$$

We can explore model behavior by writing a = 1+w, so that for example w = 0.2 represents 20% annual value growth. Roughly speaking, V[X] diverges when w significantly outstrips per-century risk *r*. When w meets or falls below per-century risk, the exponential growth model behaves somewhat similarly to the ambitious growth models considered in Section 3, such as quadratic growth. By way of illustration, here is the value of a 10% relative reduction of risk in this century across values of w,r, with DIV indicating divergence.

Table 4: Value of 10% relative risk reduction across levels of background risk and value growth under an exponential model

	r = 0.001	r = 0.01	r = 0.1	r = 0.2
w = 0.001	100.1v	0.1v	0.1v	0.1v
w = 0.01	DIV	10.1v	0.1v	0.1v
w = 0.1	DIV	DIV	1.1v	0.2v
w = 0.2	DIV	DIV	DIV	0.6v

References

Alvarez, Luis W., Alvarez, Walter, Asaro, Frank, and Michel, Helen V. 1980. "Extraterrestrial cause for the Cretaceous-Tertiary extinction." *Science* 208:1095–1180.

Aschenbrenner, Leopold. 2020. "Existential risk and growth." Global Priorities Institute Working Paper 6-2020, https://globalprioritiesinstitute.org/leopold-aschenbrennerexistential-risk-and-growth/.

Bennett, Jonathan. 1978. "On maximizing happiness." In Richard Sikora and Brian Barry (eds.), *Obligations to future generations*, 61–73. Temple University Press.

Bostrom, Nick. 2003. "Astronomical waste." *Utilitas* 15:308–14.

—. 2013. "Existential risk prevention as a global priority." *Global Policy* 4:15–31.

—. 2014. *Superintelligence*. Oxford University Press.

Bostrom, Nick and Ćirković, Milan (eds.). 2011.´ *Global catastrophic risks*. Oxford University Press.

Broome, John. 1994. “Discounting the future.” *Philosophy and Public Affairs* 23:128–56.

Ćirković, Milan. 2019. “Space colonization remains the only long-term option for humanity: A reply to Torres.” *Futures* 105:166–173.

Christensen, Peter, Gillingham, Kenneth and Nordhaus, William. 2018. “Uncertainty in forecasts of long-run economic growth.” *Proceedings of the National Academy of Sciences* 115:5409-14.

Dasgupta, Partha. 2008. “Discounting climate change.” *Journal of Risk and Uncertainty* 37:141–69.

Dasgupta, Susmita, Laplante, Benoit, Wang, Hua, and Wheeler, David. 2002. “Confronting the environmental Kuznets curve.” *Journal of Economic Perspectives* 16:147–168.

Deudney, Daniel. 2020. *Dark skies: Space expansionism, planetary geopolitics, and the ends of humanity*. Oxford University Press.

Eden, Maya and Alexandrie, Gustav. Forthcoming. “Is extinction risk mitigation uniquely cost-effective? Not in standard population models”. In Barrett, Greaves and Thorstad (eds.), *Essays on longtermism*. Oxford University Press.

Gabaix, Xavier and Laibson, David. 2022. “Myopia and discounting.” National Bureau of Economic Research Working Paper 23254, https://www.nber.org/papers/w23254.

Gardiner, Stephen. 2004. “Ethics and global climate change.” *Ethics* 114:555–600.

Geruso, Michael and Spears, Dean. Forthcoming. “With a whimper: Depopulation and longtermism.” In Barrett, Greaves and Thorstad (eds.), *Essays on longtermism*. Oxford University Press.

Gottlieb, Joseph. 2019. “Space colonization and existential risk.” *Journal of the American Philosophical Association* 5:306–320.

Greaves, Hilary. 2017. “Discounting for public policy: A survey.” *Economics and Philosophy* 33:391–439.

Greaves, Hilary and MacAskill, William. 2021. “The case for strong longtermism.” Global Priorities Institute Working Paper 5-2021, https://globalprioritiesinstitute.org/hilarygreaves-william-macaskill-the-case-for-strong-longtermism-2/.

Grossman, Gene and Krueger, Alan. 1995.“Economic growth and the environment.” *Quarterly Journal of Economics* 110:353–377.

John, Tyler and MacAskill, William. 2021. “Longtermist institutional reform.” In Natalie Cargill and Tyler John (eds.), *The long view*. FIRST.

Jones, Charles. 2016. “Life and growth.” *Journal of Political Economy* 124:539–78.

Kagan, Shelly. 1984. “Does consequentialism demand too much? Recent work on the limits of obligation.” *Philosophy and Public Affairs* 239–54.

Knutzen, Jonathan. forthcoming. “Unfinished business.” *Philosophers’ Imprint* forthcoming.

Lloyd, Harry. 2021. “Time discounting, consistency and special obligations: a defence of Robust Temporalism.” Global Priorities Institute Working Paper 11-2021, https://globalprioritiesinstitute.org/time-discounting-consistency-and-specialobligations-a-defence-of-robust-temporalism-harry-r-lloyd-yale-university/.

MacAskill, William. 2022. *What we owe the future*. Basic books.

Mogensen, Andreas. 2019. “Doomsday rings twice.” Global Priorities Institute Working Paper 1-2019.

—. 2021. “Moral demands and the far future.” *Philosophy and Phenomenological Research* 103:567–85.

—. 2022. “The only ethical argument for positive δ? Partiality and pure time preference.” *Philosophical Studies* 179:2731–50.

Monton, Bradley. 2019. “How to avoid maximizing expected utility.” *Philosophers’ Imprint* 19:1-25.

Mulgan, Tim. 2001. “How satisficers get away with murder.” *International Journal of Philosophical Studies* 9:41–46.

Musk, Elon. 2017. “Making humans a multi-planetary species.” *New Space* 5:46–61.

Ord, Toby. 2020. *The precipice*. Bloomsbury.

Parfit, Derek. 1984. *Reasons and persons*. Oxford University Press.

—. 2011. *On what matters*, volume 1. Oxford University Press.

Pettigrew, Richard. 2022. “Effective altruism, risk, and human extinction.” Global Priorities Institute Working Paper 2-2022, https://globalprioritiesinstitute.org/effectivealtruism-risk-and-human-extinction-richard-pettigrew-university-of-bristol/.

Pummer, Theron. 2016. “Whether and where to give.” *Philosophy and Public Affairs* 44:77–95.

Rees, Martin. 2003. *Our final hour*. Basic books.

Riedener, Stefan. 2021. “Existential risks from a Thomist Christian perspective.” Global Priorities Institute Working Paper 1-2021, https://globalprioritiesinstitute.org/wpcontent/uploads/Stefan-Riedener Existential-risks-from-a-Thomist-Christianperspective.pdf.

Sagan, Carl. 1997. *Pale blue dot: A vision of the human future in space*. Ballantine Books.

Sandberg, Anders and Bostrom, Nick. 2008. “Global catastrophic risks survey.” Technical Report 2008-1, Future of Humanity Institute, https://www.global-catastrophicrisks.com/docs/2008-1.pdf.

Scheffler, Samuel. 1994. *The rejection of consequentialism*. Oxford University Press.

Scheffler, Samuel and Kolodny, Niko (eds.). 2013. *Death and the afterlife*. Oxford University Press.

Schulte, Peter et al. 2010. “The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary.” *Science* 327:1214–8.

Schwartz, James. 2011. “Our moral obligation to support space exploration.” *Environmental Ethics* 33:67–88.

Singer, Peter. 1972. “Famine, affluence, and morality.” *Philosophy and Public Affairs* 1:229–43.

Smith, Nicholas. 2014. “Is evaluative compositionality a requirement of rationality?” *Mind* 123:457-502.

Sobel, David. 2007. “The impotence of the demandingness objection.” *Philosophers’ Imprint* 7:1–17.

Solow, Robert. 1956. “A contribution to the theory of economic growth.” *Quarterly Journal of Economics* 70:65–94.

Stokey, Nancy. 1998. “Are there limits to growth?” *International Economic Review* 39.

Tarsney, Christian. 2022. “The epistemic challenge to longtermism.” Global Priorities Institute Working Paper 3-2022, https://globalprioritiesinstitute.org/christian-tarsneythe-epistemic-challenge-to-longtermism/.

Tarsney, Christian and Wilkinson, Hayden. 2023. “Longtermism in an infinite world.” Global Priorities Institute Working Paper 4-2023, https://globalprioritiesinstitute.org/longtermism-in-an-infinite-world-christian-j-tarsney-and-hayden-wilkinson/.

Thompson, Dennis. 2010. “Representing future generations: Political presentism and democratic trusteeship.” *Critical Review of International Social and Political Philosophy* 13:17–37.

Tokarska, Katarzyna, Gillett, Nathan, Weaver, Andrew, Arora, Viek, and Eby, Michael. 2016. “The climate response to five trillion tonnes of carbon.” *Nature Climate Change* 6:815–55.

Torres, Phil. 2018. “Space colonization and suffering risks: Reassessing the ‘maxipok rule’.” *Futures* 100:74–85.

Trammell, Philip. 2021. “Dynamic public good provision under time preference heterogeneity: theory and applications to philanthropy.” Global Priorities Institute Working Paper 9-2021, https://globalprioritiesinstitute.org/dynamic-public-good-provisionunder-time-preference-heterogeneity-theory-and-applications-to-philanthropy-philiptrammell-global-priorities-institute-and-department-of-economics-university-ofoxford/.

Yandle, Bruce, Bhattarai, Madhusadan, and Vijayaraghavan, Maya. 2002. “The environmental Kuznets curve: A primer.” Technical report, Property and Environment Research Center.