Existential risk pessimism and the time of perils

David Thorstad (Global Priorities Institute, University of Oxford)

Global Priorities Institute | March 2022

GPI Working Paper No . 1-2022

GLOBAL PRIORITIES INSTITUTE

UNIVERSITY OF OXFORD

Existential risk pessimism and the time of perils

Abstract

Recent authors have argued that it is overwhelmingly important to mitigate existential risks: risks that threaten the survival or development of humanity. This position is often supported by pessimistically high estimates of existential risk. In this paper, I extend a model by Toby Ord and Thomas Adamczewski to do two things. First, I argue, across a range of modeling assumptions pessimism tends to hamper rather than strengthen the case for existential risk mitigation. Second, I show that pessimism is unlikely to ground the overwhelming importance of existential risk mitigation unless it is coupled with an empirical hypothesis: the time of perils hypothesis. However, I argue, the time of perils hypothesis is probably false. I conclude that existential risk pessimism may tell against the overwhelming importance of existential risk mitigation.

1 Introduction

Recent authors have argued 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, you might work to prevent the release of harmful synthetic pathogens or the premature deployment of powerful artificial intelligence systems whose values may be misaligned with our own (Bohannon 2015; Bostrom 2014; Bostrom and Ćirković 2011; Millett and Snyder-Beattie 2017; Häggström and Rhodes 2019; Ord 2020; Rees 2003).

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 provides a small probability of tremendous gain: the continued survival and development of humanity. Given the mind-boggling numbers of future human lives that may be lived, anything that we can do to ensure these lives are lived, and lived well may have astronomical value.

The case for existential risk mitigation is often supported by 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 existential risk mitigation. After all, we should usually do more to address probable threats than to address improbable threats. In this paper, I extend a model due to Toby Ord (2020; ms) and Thomas Adamczewski (ms) to do two things. First, I argue, across a range of assumptions, existential risk pessimism at best has no effect on the value of existential risk mitigation, and at worst significantly lowers the value of existential risk mitigation. Second, I use the model to explore a variety of ways in which existential risk pessimists could support the astronomical value of existential risk mitigation. I argue that the most plausible strategy is to rely on an empirical hypothesis about the future: the time of perils hypothesis on which risk is high now, but will soon fall to a low level. However, I argue that we have good reason to doubt the time of perils hypothesis. I conclude by reflecting on the strength of the case for existential risk mitigation under existential risk pessimism.

Here is the plan. Section 2 introduces the Ord-Adamczewski model of existential risk mitigation and uses the model to pose a prima facie challenge to the case for existential risk mitigation under existential risk pessimism. Section 3 evaluates a variety of ways in which pessimists could modify the Ord-Adamczewski model to support the astronomical value of existential risk mitigation and concludes that the most plausible strategy is to adopt the time of perils hypothesis. Sections 4-6 cast doubt on the case for the time of perils hypothesis. Section 7 concludes. Proofs are in the appendix.

Before beginning, I want to note an advantage of the argumentative strategy in this pa-

per. Many recent challenges to existential risk reduction proceed by questioning decision-theoretic, consequentialist or population-ethical assumptions used to build the case for existential risk reduction (Mogensen 2021; Steele forthcoming; Pettigrew manuscript; Unruh forthcoming). My argument will not do this. 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 of existential risk mitigation that may be persuasive to the pessimist herself.

2 The Simple Model

In this section, I present a Simple Model of existential risk mitigation due to Toby Ord (2020; ms) and Thomas Adamczewski (ms). 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 value of existential risk mitigation may be relatively modest. Section 3 then considers how the case for existential risk mitigation fares across modifications of the Simple Model.

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.[1] 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

[1] 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.

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 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 constrains the value of existential risk reduction to be more modest than we might have supposed. 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 case for existential risk mitigation, and that existential risk mitigation does not bear astronomical value. 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 and Adamczewski 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 recover an astronomical value of existential risk mitigation?

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 an astronomical value for existential risk mitigation. We will also see that as increasingly optimistic assumptions about value growth are considered, pessimism looms larger as a roadblock to the value of existential risk mitigation.

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.

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	$5*10^6$v

First, the value of existential 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 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 astronomical values for existential risk mitigation, even under optimistic growth models.

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 obstacles to the astronomical value of existential risk mitigation.

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 mitigation and existential risk pessimism. What else might the pessimist do to build a case for existential risk mitigation?

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 case for existential risk mitigation, 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 an astronomical value for existential risk mitigation. 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. By way of illustration, setting r to a pessimistic 20% values a 10% relative reduction 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 of existential risk mitigation.

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 increase the value of existential risk mitigation. In the next subsection, I consider the most plausible way to do this.

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 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.[2] 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. 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 $V_{\text{PERIL}} = \Sigma_{i=1}^{\infty} v(1-r)^i$ be the value of living in a world forever stuck at the perilous level of risk, and $V_{\text{SAFE}} = \Sigma_{i=1}^{\infty} v(1-r_l)^i$ 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 $Pr(\text{SAFE}) = (1-r)^N$. Then the value of the world today is a probability-weighted average of the values of the safe and perilous worlds.

[2]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).

$$V[W] = Pr(\neg \textsc{safe}) V_{\textsc{peril}} + Pr(\textsc{safe}) V_{\textsc{safe}}.$$

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 $V_{\textsc{peril}}$ 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 $V_{\textsc{safe}}$ 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:

$$V[X] = fv[1 - (1 - r)^N] + r(1 - r)^{N-1} f V_{\textsc{safe}}.$$

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, $r(1 - r)^{N-1} f V_{\textsc{safe}}$, 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 of existential risk mitigation, 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 $V_{\textsc{safe}}$ 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-

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

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.

Where does this discussion leave the pessimistic case for existential risk reduction? 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 of existential risk reduction. 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 an astronomical value for existential risk reduction. And we saw that the most likely culprit for these failures is pessimism itself.

However, we also considered 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.[3]

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).

[3]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.

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 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).

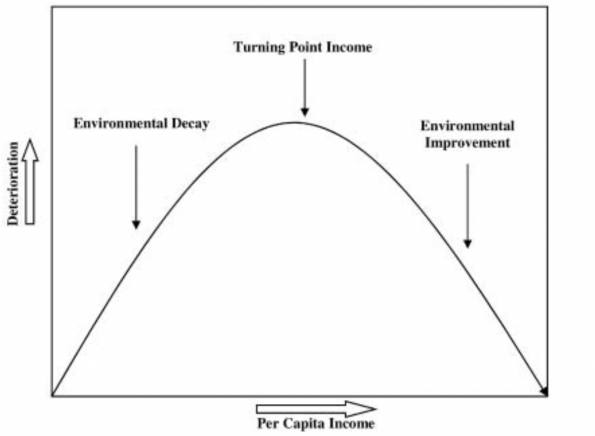
There is no doubt that institutional changes can increase the patience of political systems. For example, many of these changes have led to increased emphasis on mitigating climate risks to future generations. But the pessimist needs a mechanism by which a largely impatient group of humans could together 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,

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



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.

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.[4]

[4]These drastic scenarios might require burning the entire stock of fossil fuels on earth (Tokarska et al.

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}.

$$C_t = A_t^{\alpha} L_{ct}. \tag{1}$$

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}.

$$H_t = B_t^{\alpha} L_{ht}. \tag{2}$$

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:

$$\delta_t = \overline{\delta} C_t^{\epsilon} H_t^{-\beta}. \tag{3}$$

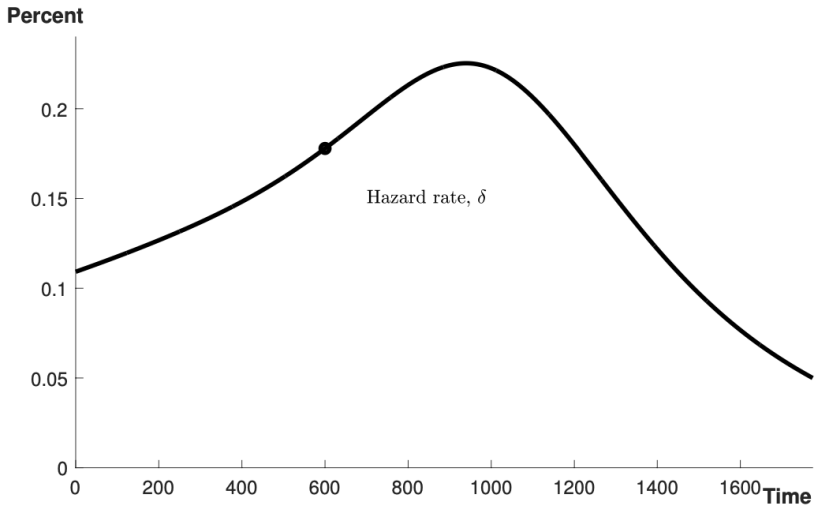
for constants $\overline{\delta}, \epsilon, \beta$.

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

2016). Even then, (Ord 2020) notes, these scenarios may well stop short of existential catastrophe.

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



(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 continuing indefinitely into the future.

I think this is probably the best argument 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 (Kaul et al. 1999). 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 under-promoted. 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.[5] 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 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,

[5]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).

and *natural risks* posed by the environment, such as asteroid impacts, super-volcanoes, 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 super-volcanoes 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 cold comfort 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 Conclusion

This paper explored the impact that pessimistic views of existential risk have on the value of existential risk mitigation. 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 suggested that the value of existential risk mitigation may be more modest than otherwise supposed.

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 an astronomical value of existential risk mitigation, 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.

Where does this 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 (Shulman and Thornley

forthcoming). 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. 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

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:

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

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

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

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:

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

So that:

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

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:

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

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

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

With X as before we have:

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

Giving:

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

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

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

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:

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

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

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

Subtracting term-wise gives:

$$
\begin{aligned}
V[X] &= V[W_X] - V[W] \\
&= \frac{v[1-(1-r)^N]}{r}[1-(1-f)r-(1-r)]+ \\
&V_{\text{SAFE}}[(1-(1-f)r)(1-r)^{N-1}-(1-r)^N] \\
&= fv[1-(1-r)^N] + (1-r)^{N-1}V_{\text{SAFE}}[1-(1-f)r-(1-r)] \\
&= fv[1-(1-r)^N] + fr(1-r)^{N-1}V_{\text{SAFE}}.
\end{aligned}
$$

References

Adamczewski, Thomas. ms. "The expected value of the long-term future." Unpublished manuscript.

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.

Bohannon, John. 2015. "Artificial intelligence: Fears of an AI pioneer." *Science* 349:252.

Bostrom, Nick. 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.

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

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.

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

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

Häggström, Olle and Rhodes, Catherine. 2019. "Guest editorial." *Foresight* 21:1–3.

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.

Kaul, Inge, Grunberg, Isabelle, and Stern, Marc (eds.). 1999. *Global public goods: International cooperation in the 21st century*. Oxford University Press.

Millett, Piers and Snyder-Beattie, Andrew. 2017. "Existential risk and cost-effective biosecurity." *Health Security* 15:373–384.

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.

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

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

—. ms. "Modelling the value of existential risk reduction." Unpublished manuscript.

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

Pettigrew, Richard. manuscript. "Effective altruism, risk, and human extinction." Manuscript.

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

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.

Schulte, Peter et al. 2010. "The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary." *Science* 327:1214–1218.

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

Shulman, Carl and Thornley, Elliott. forthcoming. "Tradeoffs between longtermism and other social metrics: Is longtermism relevant to existential risk in practice?" In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

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

Steele, Katie. forthcoming. "Risk aversion, the long term and agent-centred prerogatives." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

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

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.

Unruh, Charlotte. forthcoming. "Deontology, harm, and generational sovereignty." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

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

Existential risk pessimism and the time of perils

Abstract

Recent authors have argued that it is overwhelmingly important to mitigate existential risks: risks that threaten the survival or development of humanity. This position is often supported by pessimistically high estimates of existential risk. In this paper, I extend a model by Toby Ord and Thomas Adamczewski to do two things. First, I argue, across a range of modeling assumptions pessimism tends to hamper rather than strengthen the case for existential risk mitigation. Second, I show that pessimism is unlikely to ground the overwhelming importance of existential risk mitigation unless it is coupled with an empirical hypothesis: the time of perils hypothesis. However, I argue, the time of perils hypothesis is probably false. I conclude that existential risk pessimism may tell against the overwhelming importance of existential risk mitigation.

1 Introduction

Recent authors have argued 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, you might work to prevent the release of harmful synthetic pathogens or the premature deployment of powerful artificial intelligence systems whose values may be misaligned with our own (Bohannon 2015; Bostrom 2014; Bostrom and Ćirković 2011; Millett and Snyder-Beattie 2017; Häggström and Rhodes 2019; Ord 2020; Rees 2003).

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 provides a small probability of tremendous gain: the continued survival and development of humanity. Given the mind-boggling numbers of future human lives that may be lived, anything that we can do to ensure these lives are lived, and lived well may have astronomical value.

The case for existential risk mitigation is often supported by 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 existential risk mitigation. After all, we should usually do more to address probable threats than to address improbable threats. In this paper, I extend a model due to Toby Ord (2020; ms) and Thomas Adamczewski (ms) to do two things. First, I argue, across a range of assumptions, existential risk pessimism at best has no effect on the value of existential risk mitigation, and at worst significantly lowers the value of existential risk mitigation. Second, I use the model to explore a variety of ways in which existential risk pessimists could support the astronomical value of existential risk mitigation. I argue that the most plausible strategy is to rely on an empirical hypothesis about the future: the time of perils hypothesis on which risk is high now, but will soon fall to a low level. However, I argue that we have good reason to doubt the time of perils hypothesis. I conclude by reflecting on the strength of the case for existential risk mitigation under existential risk pessimism.

Here is the plan. Section 2 introduces the Ord-Adamczewski model of existential risk mitigation and uses the model to pose a prima facie challenge to the case for existential risk mitigation under existential risk pessimism. Section 3 evaluates a variety of ways in which pessimists could modify the Ord-Adamczewski model to support the astronomical value of existential risk mitigation and concludes that the most plausible strategy is to adopt the time of perils hypothesis. Sections 4-6 cast doubt on the case for the time of perils hypothesis. Section 7 concludes. Proofs are in the appendix.

Before beginning, I want to note an advantage of the argumentative strategy in this pa-

per. Many recent challenges to existential risk reduction proceed by questioning decision-theoretic, consequentialist or population-ethical assumptions used to build the case for existential risk reduction (Mogensen 2021; Steele forthcoming; Pettigrew manuscript; Unruh forthcoming). My argument will not do this. 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 of existential risk mitigation that may be persuasive to the pessimist herself.

2 The Simple Model

In this section, I present a Simple Model of existential risk mitigation due to Toby Ord (2020; ms) and Thomas Adamczewski (ms). 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 value of existential risk mitigation may be relatively modest. Section 3 then considers how the case for existential risk mitigation fares across modifications of the Simple Model.

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.[1] 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

[1] 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.

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 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 constrains the value of existential risk reduction to be more modest than we might have supposed. 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 case for existential risk mitigation, and that existential risk mitigation does not bear astronomical value. 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 and Adamczewski 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 recover an astronomical value of existential risk mitigation?

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 an astronomical value for existential risk mitigation. We will also see that as increasingly optimistic assumptions about value growth are considered, pessimism looms larger as a roadblock to the value of existential risk mitigation.

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.

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	$5*10^6$v

First, the value of existential 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 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 astronomical values for existential risk mitigation, even under optimistic growth models.

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 obstacles to the astronomical value of existential risk mitigation.

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 mitigation and existential risk pessimism. What else might the pessimist do to build a case for existential risk mitigation?

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 case for existential risk mitigation, 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 an astronomical value for existential risk mitigation. 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. By way of illustration, setting r to a pessimistic 20% values a 10% relative reduction 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 of existential risk mitigation.

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 increase the value of existential risk mitigation. In the next subsection, I consider the most plausible way to do this.

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 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.[2] 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. 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 $V_{\text{PERIL}} = \Sigma_{i=1}^{\infty} v(1-r)^i$ be the value of living in a world forever stuck at the perilous level of risk, and $V_{\text{SAFE}} = \Sigma_{i=1}^{\infty} v(1-r_l)^i$ 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 $Pr(\text{SAFE}) = (1-r)^N$. Then the value of the world today is a probability-weighted average of the values of the safe and perilous worlds.

[2]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).

$$V[W] = Pr(\neg \textsc{safe}) V_{\textsc{peril}} + Pr(\textsc{safe}) V_{\textsc{safe}}.$$

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 $V_{\textsc{peril}}$ 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 $V_{\textsc{safe}}$ 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:

$$V[X] = fv[1 - (1 - r)^N] + r(1 - r)^{N-1} f V_{\textsc{safe}}.$$

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, $r(1 - r)^{N-1} f V_{\textsc{safe}}$, 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 of existential risk mitigation, 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 $V_{\textsc{safe}}$ 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-

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

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.

Where does this discussion leave the pessimistic case for existential risk reduction? 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 of existential risk reduction. 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 an astronomical value for existential risk reduction. And we saw that the most likely culprit for these failures is pessimism itself.

However, we also considered 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.[3]

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).

[3]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.

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 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).

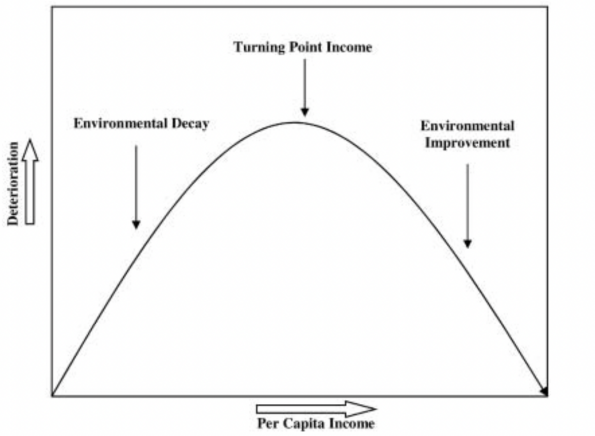
There is no doubt that institutional changes can increase the patience of political systems. For example, many of these changes have led to increased emphasis on mitigating climate risks to future generations. But the pessimist needs a mechanism by which a largely impatient group of humans could together 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,

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



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.

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.[4]

[4]These drastic scenarios might require burning the entire stock of fossil fuels on earth (Tokarska et al.

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}.

$$C_t = A_t^{\alpha} L_{ct}. \tag{1}$$

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}.

$$H_t = B_t^{\alpha} L_{ht}. \tag{2}$$

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:

$$\delta_t = \overline{\delta} C_t^{\epsilon} H_t^{-\beta}. \tag{3}$$

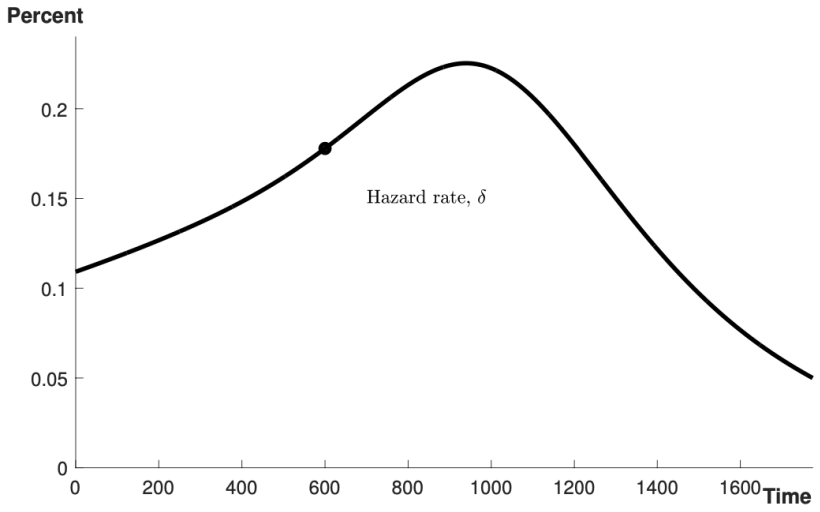
for constants $\overline{\delta}, \epsilon, \beta$.

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

2016). Even then, (Ord 2020) notes, these scenarios may well stop short of existential catastrophe.

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



(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 continuing indefinitely into the future.

I think this is probably the best argument 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 (Kaul et al. 1999). 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 under-promoted. 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.[5] 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 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,

[5]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).

and *natural risks* posed by the environment, such as asteroid impacts, super-volcanoes, 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 super-volcanoes 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 cold comfort 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 Conclusion

This paper explored the impact that pessimistic views of existential risk have on the value of existential risk mitigation. 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 suggested that the value of existential risk mitigation may be more modest than otherwise supposed.

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 an astronomical value of existential risk mitigation, 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.

Where does this 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 (Shulman and Thornley

forthcoming). 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. 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

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:

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

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

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

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:

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

So that:

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

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:

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

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

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

With X as before we have:

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

Giving:

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

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

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

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:

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

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

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

Subtracting term-wise gives:

$$
\begin{aligned}
V[X] &= V[W_X] - V[W] \\
&= \frac{v[1-(1-r)^N]}{r}[1-(1-f)r-(1-r)]+ \\
&V_{\text{SAFE}}[(1-(1-f)r)(1-r)^{N-1}-(1-r)^N] \\
&= fv[1-(1-r)^N] + (1-r)^{N-1}V_{\text{SAFE}}[1-(1-f)r-(1-r)] \\
&= fv[1-(1-r)^N] + fr(1-r)^{N-1}V_{\text{SAFE}}.
\end{aligned}
$$

References

Adamczewski, Thomas. ms. "The expected value of the long-term future." Unpublished manuscript.

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.

Bohannon, John. 2015. "Artificial intelligence: Fears of an AI pioneer." *Science* 349:252.

Bostrom, Nick. 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.

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

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.

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

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

Häggström, Olle and Rhodes, Catherine. 2019. "Guest editorial." *Foresight* 21:1–3.

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.

Kaul, Inge, Grunberg, Isabelle, and Stern, Marc (eds.). 1999. *Global public goods: International cooperation in the 21st century*. Oxford University Press.

Millett, Piers and Snyder-Beattie, Andrew. 2017. "Existential risk and cost-effective biosecurity." *Health Security* 15:373–384.

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.

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

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

—. ms. "Modelling the value of existential risk reduction." Unpublished manuscript.

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

Pettigrew, Richard. manuscript. "Effective altruism, risk, and human extinction." Manuscript.

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

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.

Schulte, Peter et al. 2010. "The Chicxulub asteroid impact and mass extinction at the Cretaceous-Paleogene boundary." *Science* 327:1214–1218.

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

Shulman, Carl and Thornley, Elliott. forthcoming. "Tradeoffs between longtermism and other social metrics: Is longtermism relevant to existential risk in practice?" In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

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

Steele, Katie. forthcoming. "Risk aversion, the long term and agent-centred prerogatives." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

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

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.

Unruh, Charlotte. forthcoming. "Deontology, harm, and generational sovereignty." In Jacob Barrett, Hilary Greaves, and David Thorstad (eds.), *Longtermism*. Oxford University Press.

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