RETHINK PRIORITIES

Technical Report: The Value of Existential Risk Mitigation

Worldview Investigations Team

23 October - 2023

AUTHOR

Arvo Muñoz Morán – Quantitative Researcher

Abstract

This report investigates the value of existential risk mitigation by extending the model developed by Ord, Thorstad and Adamczewski. The work here uses more realistic assumptions, like sophisticated risk structures, variable persistence and new cases of value growth. By enriching the base model, we are able to perform sensitivity analyses and can better evaluate when existential risk mitigation should, in expectation, be a global priority.

Keywords: counterfactual value; extinction risk; existential risk mitigation; x-risk; sensitivity analysis; expected value of the future; persistence; longtermism; long-term effects; Great Filters Hypothesis; Time of Perils Hypothesis.

This report is accompanied by an interactive Jupyter Notebook. The post version of the abridged report also features an executive summary. This is the full report.

This report is a part of Rethink Priorities' Worldview Investigations Team's CURVE Sequence: "Causes and Uncertainty: Rethinking Value in Expectation." The aim of this sequence is twofold: first, to consider alternatives to expected value maximisation for cause prioritisation; second, to evaluate the claim that a commitment to expected value maximisation robustly supports the conclusion that we ought to prioritise existential risk mitigation over all else.

Contents

I Modelling the Value of Risk Mitigation 6

1 Introduction 7

1.1 OAT Limitations . . . 8

1.2 Key Research Questions . . . 8

2 Generalised Model: Arbitrary Risk Profile 10

2.1 Value . . . 11

2.1.1 V0 Constant Value . . . 11

2.1.2 V1 Linear Value . . . 12

2.1.3 V2 Quadratic Value . . . 12

2.1.4 V3 Cubic Value . . . 12

2.1.5 V4 Logistic Value . . . 12

2.1.6 Value Cases Summary . . . 13

2.2 Persistence . . . 13

2.2.1 A Concrete Example . . . 14

2.3 The Rest of this Report . . . 16

3 Great Filters and the Time of Perils Hypothesis 18

4 Decaying Risk 19

4.0.1 Risk Cases Summary . . . 19

II Results 20

5 Convergence 21

5.0.1 Constant $v_t = v_c$ . . . 22
5.0.2 Linear $v_t = v_c\, i$ . . . 22
5.0.3 Quadratic $v_t = v_c\, i^2$ . . . 23
5.0.4 Cubic $v_t = v_c\, i^3$ . . . 23
5.0.5 Polynomial $v_t = v_c\, i^n$ . . . 23
5.0.6 Polynomial under Adjusted Sequences . . . 23
5.0.7 Logistic Value . . . 23

6 Closed Form Solutions **24**
6.1 Constant Risk . . . 25

7 The Expected Value of Mitigating Risk Visualised **27**

8 Concluding Remarks **33**

9 References **34**

10 Acknowledgements **36**

11 Century to Year Adjustments **37**

12 Other Closed Forms **38**
12.1 Constant risk with polynomial value of order n . . . 38
12.2 Constant Risk closed forms given yearly adjustments . . . 39
12.2.1 Constant . . . 39
12.2.2 Linear . . . 39
12.2.3 Quadratic . . . 40
12.2.4 Cubic . . . 41

Part I

Modelling the Value of Risk Mitigation

1 Introduction

Consider a catastrophe that permanently ends human civilisation.[1] You might find it plausible that any efforts to reduce the risk of such a catastrophe are of enormous value. You might also be inclined to think that the value is particularly high if the risks are high also. After all, in most contexts, the bigger the risk of something bad happening, the less it can be safely ignored. In other words, you might believe that it is of astronomical importance to mitigate these 'extinction risks' because the stakes are very large and because the probability of these catastrophic scenarios is uncomfortably high. Existing work by Ord, Adamczewski and Thorstad (hereon 'OAT') argues that this last sentence is questionable: in the context of an extinction catastrophe, the higher we think the risk is, the less we should value efforts that mitigate that risk.[2]

Our initial intuitions are not always a good guide for how we should think about estimating the value of extinction risk mitigation. Indeed, the unexpected tensions between high pessimism about the risk we face and whether risk mitigation is of astronomical value, are a good example of this.[3] Similarly, simplified attempts and heuristics used to estimate the cost effectiveness of risk reduced –– such as 1, 2, 3, 4, 5 –– turn out to only be appropriate in a handful of very restricted scenarios (usually where value and risk are constant in all the periods), and they otherwise mischaracterise the value of extinction risk mitigation.

If we want to evaluate the general merits of interventions that seek to safeguard humanity's future, we need a systematic way to estimate the value of mitigating extinction risk. The current frameworks help us understand which scenarios might lead to astronomical value. However, they have several limitations that make it difficult, or sometimes impossible, to comment meaningfully on the amount of good that mitigating risk in the next few decades could achieve. This report builds on the existing models and provides tools to estimate the value of mitigating risk in more realistic settings.

The Base Model

As a first attempt to provide a more rigorous analysis, existing work presents a stylised model to assess the value of extinction risk mitigation given the following assumptions:

A1 Each century of human existence has some constant value.

A2 Humans face a constant level of per-century extinction risk.

A3 No value will be realised after an extinction catastrophe.

[1] Previous work has referred to such a risk as 'existential risk'. But this is a misnomer. Existential risk is technically broader and it encompasses another case: the risk of an event that drastically and permanently curtails the potential of humanity. For the rest of this report we characterise the risk as that of extinction where previous work has used 'existential'.

[2] The reasoning goes that if there is always a high level of background risk to humanity, then we should expect to go extinct soon anyway, which means the importance of avoiding any one particular risk is not as valuable as it may seem.

[3] In particular, Thorstad explores how, in this model, extinction risk pessimism fails to support and sometimes hinders the thesis that extinction risk mitigation is of astronomical value.

A4 Risk is reduced by a fraction.

A5 Risk is only reduced this century.

A6 Centuries are the shortest time units.

The model is clearly oversimplified, and, indeed, previous work has partially relaxed a subset of these six assumptions.[4] However, there are still several limitations present in those frameworks.

1.1 OAT Limitations

Some of the main limitations of the previous work include:

- The current models lack the necessary resolution to yield results that are relevant for, or incorporate observations from, key issues like near-term AI timelines. The models cannot presently handle anything requiring shorter time units than centuries.
- The duration of a mitigation action's effects affects its overall value. However, OAT has not explored how varying the duration of these effects may impact the model.[5]
- There are many possible scenarios (i.e., combinations of risk and value trajectories), and OAT has explored very few of these. Given our large uncertainty in this area, it is a priority to have a clear picture of how the value compares in each case. This will provide the necessary tools for future work that assigns credences to each scenario to arrive at better-informed expected value judgements.
- There are currently no versatile frameworks that can calculate the expected value of mitigating risk, for a given set of idiosyncratic beliefs about risk and value trajectories.
- As time goes to infinity, the expected value of existential risk mitigation could, in principle, be infinite; making most scenario comparisons redundant in those cases. There has been no formal discussion of the convergence of the value of extinction risk mitigation for all of the main scenarios.

1.2 Key Research Questions

The present report aims to tackle all of the above limitations. With that in mind, the key guiding questions are:

1. When is the value – of the future and of risk mitigation – particularly large and when is it not?

[4] For example, Thorstad relaxes each of the A1, A4 and A5 assumptions.

[5] The models thus far centred around mitigating risk for one century only. Thorstad comments on one additional case: when risk is permanently mitigated, calling it 'global risk reduction'.

2. What is the Great Filter Hypothesis, how does it relate to the Time of Perils and what is the impact of adding great filters on the value of risk mitigation?
3. What are the qualitative pictures of the expected value of the world – and thus of mitigation efforts – given different risk structures (e.g. linear, Time of Perils, Great Filters, decaying) and value growth cases (e.g. linear, quadratic, cubic, logistic)?
4. How does the value of mitigation efforts depend on their persistence?

The main ambition here is to develop a generalised version of the toy model that relaxes all assumptions above, except for A3, no value after extinction, and A4, fractional risk reduction.[6] [7] By relaxing A1 and A2 – that the value and risk are constant – we are able to introduce a framework that can accommodate more complex risk structures and sophisticated value trajectories. We also depart from existing analyses by relaxing A6: here, years are the shortest time unit. Moreover, by also relaxing A5, the model now has tools to observe persistence of mitigation effects lasting less (or more) than one century and can meaningfully comment on the near-term value of extinction risk mitigation. Using this generalised framework, we can systematically assess the value of risk mitigation under various combinations of assumptions.

[6] We leave A4 untouched because it introduces diminishing returns in risk reduction (see more the details Adamczewski discusses), which we find realistic.

[7] A3 is a core assumption in the extended and simplified versions of this model. Relaxing it would amount to changing the approach completely.

2 Generalised Model: Arbitrary Risk Profile

Let us consider the expected value of a world w that faces an existential risk r_t at time t. This is best observed with a picture.

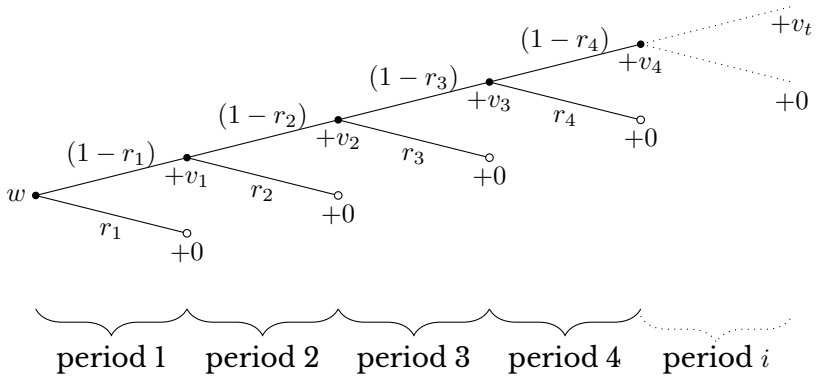


Figure 1: The Value of a World Facing Existential Risk

At each period t the world ends with probability r_t and all possible future value is reduced to zero. On the other hand, with probability $(1-r_t)$, the world progresses to the next period and achieves value v_t, which is added to the total pool of value it had accrued. Figure 1 summarises all of this. The expected value is the value of each branch weighted by the probability of reaching that value. That is

$$\mathbb{E}(w) = r_1 \cdot 0 + (1-r_1)v_1 + (1-r_1)r_2 \cdot 0 + (1-r_1)(1-r_2)v_2 + (1-r_1)(1-r_2)r_3 \cdot 0 + (1-r_1)(1-r_2)(1-r_3)v_3 + ...$$

In other words, the expected value of this world is

$$\begin{aligned}\mathbb{E}(w) &= (1-r_1)v_1 + (1-r_1)(1-r_2)v_2 + (1-r_1)(1-r_2)(1-r_3)v_3 + ... \\ &= \sum_{t=1}^{T}\left[\left(\prod_{j=1}^{t}(1-r_j)\right)v_t\right]. \end{aligned} \tag{1}$$

where the maximum number of periods T is the age of the universe when it ends, and $T \to \infty$ when we assume an infinite universe. We do not impose that $T \to \infty$ or otherwise to give the flexibility to consider cases where there is some known, exogenous, end to the universe. Throughout this document, the length of a period will equal one year. However, the results are not tied to any particular interpretation of period length.[8]

[8] That said, the risk and value trajectories usually need adjusting when considering a different time unit. For more details see section 11.

Now consider a risk mitigation action M which reduces the original risk sequence from r to r', where, for some t, $r'_t = (1-f)r_t$ and $f \in (0,1)$ is the fraction of the risk that is successfully mitigated.[9] [10] What value have we added by performing action M? In the most basic sense, we have changed the expected value of the future by

$$\mathbb{E}(M) = \mathbb{E}(w') - \mathbb{E}(w),$$

where our action modified the original risk from r in world w to r' in w'.[11] More generally, we could allow $f \leq 0$, which would amount to increasing the risk and M would produce negative value (or none at all if $f = 0$). For example, $f < 0$ if M made a nuclear war more likely by contributing to political instability. For the rest of the report we focus on non-negative value.

2.1 Value

Denote v as $v_1, v_2, v_3, v_4, ...$ as the sequence of values that the world will follow, conditional on the world existing at time t. Estimating this sequence is no trivial undertaking. There is large uncertainty in this area and considerable research is needed for us to insert reasonable values into the sequence v. Given this uncertainty, a promising approach is to develop a more flexible framework, i.e. the generalised model above and its accompanying code in the Jupyter Notebook, that is versatile enough to handle a wide range of cases. Next, we will investigate several possible paths for value growth, in particular: constant, linear, quadratic, cubic and logistic.

V0 Constant

V1 Linear

V2 Quadratic

V3 Cubic

V4 Logistic

2.1.1 V0 Constant Value

As a benchmark, we will often assume that the value available at each period is always v_c, that is $v_t = v_c, \forall i \leq T$. Then, Equation 1 becomes v_c times the expected number of periods survived. That is,

[9] In its most general form, r' could be any new risk vector that M has brought about. All there is left to evaluate the value of the action is to compute $\mathbb{E}(w') - \mathbb{E}(w)$.

[10] Alternatively, an altruistic intervention could seek to improve the future by positively influencing the value trajectory; that is, by bringing about a better v' rather than a new r'. Such actions, deserve a separate analysis.

[11] So far we have been writing $\mathbb{E}(w)$ to abbreviate $\mathbb{E}(w(r, v, T))$, where r, v and T are, respectively, the risk vector (sometimes termed 'risk profile'), the value vector and the maximum number of periods in our universe, which could be infinite. Note that a different class of interventions might focus on increasing the value of the world from $v = (v_1, v_2, ...)$ to $v' = (v_1, v_2, ...)$, which would also result in negative value according to $\mathbb{E}(w) - \mathbb{E}(w')$. Exploring these is not within the scope of this report.

$$\mathbb{E}(w) = v_c \sum_{t=1}^{T} \left[\left(\prod_{j=1}^{t} (1 - r_j) \right) \right]. \tag{2}$$

2.1.2 V1 Linear Value

If instead we assumed that the value available at each period is $iv_c, \forall i \leq T$, then

$$\mathbb{E}(w) = v_c \sum_{t=1}^{T} \left[\left(\prod_{j=1}^{t} (1 - r_j) \right) i \right]. \tag{3}$$

2.1.3 V2 Quadratic Value

The quadratic case assumes that the value available at each period is $i^2 v_c, \forall i \leq T$, so

$$\mathbb{E}(w) = v_c \sum_{t=1}^{T} \left[\left(\prod_{j=1}^{t} (1 - r_j) \right) i^2 \right]. \tag{4}$$

2.1.4 V3 Cubic Value

Similarly, cubic is $i^3 v_c, \forall i \leq T$, so

$$\mathbb{E}(w) = v_c \sum_{t=1}^{T} \left[\left(\prod_{j=1}^{t} (1 - r_j) \right) i^3 \right]. \tag{5}$$

This concludes the 'standard' cases for value trajectories.[12]

2.1.5 V4 Logistic Value

Logistic can be thought of as 'exponential with a value cap', a model that has special economic relevance. In this case we have $v_t = v_c(c\left(1 - e^{-\left(\frac{i}{a}\right)^b}\right) + 1)$, where $a \in \mathbb{Z}^+$ represents the number of periods by which the value attains more than half its cap (and

[12]The previous cases V0, V1 and V2 are 'standard' in that they follow from conventions explicitly used in OAT, and V3 and V4 are the natural extensions of them. However, since we are replacing centuries with years in this generalised model, some adjustments should be made in how these value cases are implemented. See all the details in section 11.

usually about two thirds of it), $b > 2$ affects how many periods pass before the value explodes, and c is the value cap.[13] Thus,

$$\mathbb{E}(w) = v_c \sum_{t=1}^{T} \left[\left(\prod_{j=1}^{t} (1 - r_j) \right) \left(c \left(1 - e^{-\left(\frac{t}{a}\right)^b} \right) + 1 \right) \right]. \tag{6}$$

$$v_t = \frac{c}{1 + \frac{c-s}{s} e^{-\gamma t}} \tag{7}$$

Where:

- c is the carrying capacity (the maximum value the v_t can reach).

- s is the initial value at $t = 0$. This is set so that $v_1 = 1$, which is v_c normalised.

- γ is the growth rate.

2.1.6 Value Cases Summary

Here is a table summary of the main value cases this report will investigate.[14] When the time unit is years instead of centuries, the value is adjusted to reflect this (see the full report here for the details). Cubic has previously been adopted for modelling interplanetary expansion. Logistic can be thought of as 'exponential with a value cap', a model that has special economic relevance.[15]

Table 1: Summary of v_t Cases

	Constant	**Linear**	**Quadratic**	**Cubic**	**Logistic**
v_t	v_c	tv_c	$t^2 v_c$	$t^3 v_c$	$\frac{c}{1+\frac{c-s}{s}e^{-\gamma t}}$

Here is a visual summary.

2.2 Persistence

Extinction risk mitigation actions could have effects that last different amounts of time. We may have reasons to believe that an action will reduce risks only for a few years; for

[13] a is adjusted depending on the scale in section 11. A higher b means more delay. For example, $a = 500$ paired with $b = 2$ in the context of years, gives a few decades before an explosion, perhaps because of transformative AI. For an improved visualisation, Figure 2 uses $c = 500^3, a = 100/\alpha, b = 9.$.

[14] Here: v_t is the value at time t, c is the cap value the v_t can reach and s is the starting value at $t = 0$. v_c is a constant, normalised to 1 in all the simulations. More generally, we interpret v_c as one year of value in 2023, which in human terms is roughly 8 billion people enjoying life at an average of 0.85QALYs each.

[15] Other work, has considered exponential without a cap. There seem to be good reasons to posit a cap, however high, like the physical limits on how much matter is accessible to humans in our expanding universe.

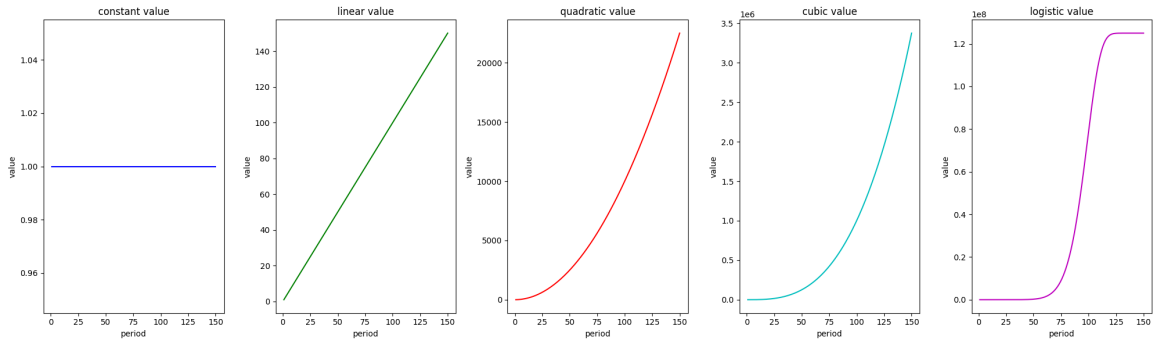


Figure 2: Value Cases

example, passing a bill that restricts AI compute which is expected to be overturned after the next election cycle in 5 years. Other actions could last longer; for example, a shield in space that physically protects Earth from asteroid impact could be effective for thousands of years. Or, in the extreme case, an action could reduce extinction risk forever. In this report, we refer to the length of the mitigating effect of an action as its *persistence*.

Persistence is key in evaluating the value of an action M. In the Ord model, the persistence of M has been assumed to be of exactly one period (which equals one century in that setting). Thorstad proceeds with the same assumption and briefly considers the permanent case as well. Because persistence plays such an important role, we developed a more flexible framework where we allow persistence P to be anything between one period and permanently reducing risk, i.e. $P \in \mathbb{Z}^+$.

An investigation of persistence likely deserves a report of its own, both for a theoretical and empirical treatment of the issue. For now we will assume that M mitigates risk for P periods, without delay. We illustrate how results differ by presenting five representative cases: $P = 1, 5, 50, 500, 2000$.

So, for example, if we had a risk profile of $r = (0.5, 0.5, 0.2, 0.4, 0.1, 0.2, ...)$ and M acts at the first period with persistence $P = 3$ and an efficacy of $f = 0.5$, halving the risk, the profile then becomes: $r' = (0.25, 0.25, 0.1, 0.4, 0.1, 0.2, ...)$.

2.2.1 A Concrete Example

There are too many cases for us to explicitly consider each one in the exposition of this report. Instead, they are systematically solved for and implemented in the code; so the user can see the results for any one desired scenario. However, it is pedagogically valuable to explicitly discuss one of these cases here.

$\mathbb{E}(w)$ under Constant Risk

If the risk were always $r_c \in (0, 1)$, Equation 1 becomes

$$E(w) = \sum_{t=1}^{T} \left[(1 - r_c)^t \, v_t \right]. \tag{8}$$

$\mathbb{E}(w)$ under Constant Value and Risk

If both the risk and the value are constant, we obtain Ord's model

$$E(w) = v_c \sum_{t=1}^{T} \left[(1 - r_c)^t \right] = \frac{v_c}{r_c}. \tag{9}$$

The value of mitigation when risk takes two values

Take the above case of constant risk and value. Counterfactual credit and Equation 2 tell us that

$$\begin{aligned}
\mathbb{E}(M) =& E(w') - E(w) \\
=& v_c \sum_{t=1}^{T} \left[\prod_{j=1}^{t} (1 - r'_j) \right] - v_c \sum_{t=1}^{T} \left[\prod_{j=1}^{t} (1 - r_j) \right] \\
=& v_c \left(\sum_{t=1}^{T} \left[\prod_{j=1}^{t} (1 - r'_j) \right] - \sum_{t=1}^{T} \left[\prod_{j=1}^{t} (1 - r_j) \right] \right).
\end{aligned}$$

Suppose that performing M halves the risk with a 5-year persistence. Let us also add some complexity to the risk structure, so it takes two constant values. Suppose that there is a 0.2229% annual risk, which approximates a one in five chance of surviving the end of the century, under the assumption that it remains constant for the next 100 years.[16] Suppose that, for no particular reason, the annual risk after those 100 years is 0.01%.[17] That is $r = (0.2229\%, 0.2229\%, ..., 0.2229\%, 0.01\%, 0.01\%, ...)$. Suppose, for this exercise, that this universe lasts 10,000 years. [18] We also normalise the value of each year to $v_c = 1$. What is the value of performing M?

Without performing M, the expected value of the world's future is

[16] The probability of dying each year that is congruent with a 0.2 probability of dying over 100 years is approximately 0.00222894771 or 0.2229%. To see why, consider the following binary outcomes model. Let p be the probability of dying in a given year. The implied probability of surviving for one year is $1 - p$. The probability of surviving for 100 years consecutively would be $(1 - p)^{100}$. Given that there's a 0.2 probability of dying over 100 years, the probability of surviving the entire 100 years is $1 - 0.2 = 0.8$. Thus, $(1 - p)^{100} = 0.8$.

[17] Which is congruent with a $(1 - 0.0001)^{100} \approx 0.99004933869$ probability of surviving each century.

[18] Numerical approximations of the expected value of M converge in this setting for large T so the universe could be thought of as infinite. See section 5 for a discussion of convergence.

$$
\begin{aligned}
\mathbb{E}(w) =& v_c \sum_{t=1}^{10000} \left[\prod_{j=1}^{t} (1 - r_j) \right] \\
=& \sum_{t=1}^{100} \left[\prod_{j=1}^{t} (1 - r_j) \right] + \sum_{t=101}^{10000} \left[\prod_{j=1}^{t} (1 - r_j) \right] \\
=& \sum_{t=1}^{100} (1 - 0.002229)^t + (1 - 0.002229)^{100} \sum_{t=1}^{9900} (1 - 0.0001)^t \\
\approx& 5116.53273619555,
\end{aligned}
$$

where the last line uses the script in the companion notebook. When performing M, the expected value of the world's future is

$$
\begin{aligned}
\mathbb{E}(w') =& v_c \sum_{t=1}^{10000} \left[\prod_{j=1}^{t} (1 - r'_j) \right] \\
=& \underbrace{\sum_{t=1}^{5} \left[\prod_{j=1}^{t} (1 - f r_j) \right]}_{\text{Risk is mitigated for 5 periods}} + \sum_{t=6}^{100} \left[\prod_{j=1}^{t} (1 - r'_j) \right] + \sum_{t=101}^{10000} \left[\prod_{j=1}^{t} (1 - r'_j) \right] \\
=& \sum_{t=1}^{5} \beta^t + \beta^5 \sum_{t=1}^{95} (1 - 0.002229)^t + \beta^5 (1 - 0.002229)^{95} \sum_{t=1}^{9900} (1 - 0.0001)^t \\
\approx& 5145.161060930257 \,,
\end{aligned}
$$

where $\beta \equiv (1 - \frac{0.002229}{2})$. Thus, the value of performing M is

$$
\mathbb{E}(M) = \mathbb{E}(w') - \mathbb{E}(w) \approx 28.6.
$$

It is worth roughly 28.6 years of a world like ours to perform M under these assumptions.

2.3 The Rest of this Report

So far, we have thought about risk in the abstract. Indeed, what we have outlined is enough for us to evaluate any arbitrary risk and value structure that we may want to test. See the Jupyter Notebook to try this yourself.

However, there are specific risk structures that we might be especially interested in evaluating. We might be inclined to believe certain stories about risk; for example, that it will systematically decline (like in section 4). Alternatively, we might want to pay heed to the commonly held view that humanity is living in a particularly risky

period now, but will reach a low-risk future if it overcomes the present challenges. The concrete example above is an instance of this, assuming constant value. Thorstad states this view, termed the 'Time of Perils' hypothesis, as:

> **(ToP)** Existential risk is and will remain high for several years, but drop to a low level if humanity survives this Time of Perils.[19]

We explore this type of risk structure next.

3 Great Filters and the Time of Perils Hypothesis

Humanity is potentially facing unprecedented threats from nuclear weapons, engineered pandemics and advanced artificial intelligence, among others. It may be that we are living in perilous times. If we do well, we might escape these dangers. But who's to say that there will be no comparable challenges in the future? The perilous times might return.

The reasoning above introduces the notion of great filters: hurdles that our civilisation must pass to ensure its long-term longevity (Hanson, 1998).[20] Specific details as to what these filters might be are beyond this work. But if AI is the first filter, we could easily imagine future ones such as escaping our dying sun or meeting powerful and unfriendly alien life. The great filter hypothesis tells us:

(GFH) Humanity will face one or more great filters, during which extinction risk will be unusually
Otherwise, the risk will be low.

It follows that, by construction, the Time of Perils hypothesis is the one filter version of GFH. For the purposes of this report, let us consider a stylised model of GFH where:

1. There are $F \in \mathbb{Z}^+$ filters (e.g. $F = 2$).
2. There are $2F$ 'eras', sets of periods within which risk is constant. Filters are high-risk eras.
3. Filters and low-risk eras alternate, starting with a filter.
4. The length of each era is given by $\ell = (\ell_1, \ell_2, ..., \ell_{2F})$.[21]
5. At each era i, humanity faces a per-period constant risk g_i, and g denotes the vector $(g_1, g_2, ...g_{2F})$.

For example, suppose that we had $F = 2$, such that there are two filters, with two lower-risk eras of lower risk after each of them. Suppose that $g = (r_1, r_{low}, r_2, r_{low})$, $\ell = (100, 500, 100, 10^{100})$ and that value is constant. From this we could write the expected value of such a world as

$$\mathbb{E}(w) = \underbrace{\sum_{t=1}^{100} v_c(1-r_1)^t}_{\text{First filter}} + \underbrace{(1-r_1)^{100} \sum_{t=1}^{500} v_c(1-r_{low})^t}_{\text{Low-risk era}} \tag{10}$$
$$+ \underbrace{(1-r_1)^{100}(1-r_{low})^{500} \sum_{t=1}^{100} v_c(1-r_2)^t}_{\text{Second filter}} + \underbrace{(1-r_1)^{100}(1-r_{low})^{500}(1-r_2)^{100} \sum_{t=1}^{10^{100}} v_c(1-r_{low})^t}_{\text{Low-risk era}}$$

[20] An excellent informal introduction to great filters can be found here.
[21] $\ell_i \in \mathbb{Z}^+$ for all $i < 2F$. In the infinitely-long universe case $\ell_{2F} \to \infty$.

4 Decaying Risk

Optimistically, we could live in a world where humanity is progressively getting better at surviving. One way of modelling this is with decreasing risk, and in particular, we can specify an exponentially decreasing function; where $r_0 \in (0, 1)$ is the starting risk, $\lambda \in (0, 1)$ is the decay rate, t is the period, $r(t)$ is the risk in period t and $r_\infty \in [0, 1)$ is the risk as $t \to \infty$. For the first few periods the sequence is: $r_0, \quad r_0 e^{-\lambda}, \quad r_0 e^{-2\lambda}, \quad r_0 e^{-3\lambda}, \quad \ldots$ More generally,

$$r(t) = r_0 \cdot e^{-\lambda t} + r_\infty.$$

4.0.1 Risk Cases Summary

A graph summarising the main cases of interest can be found below.

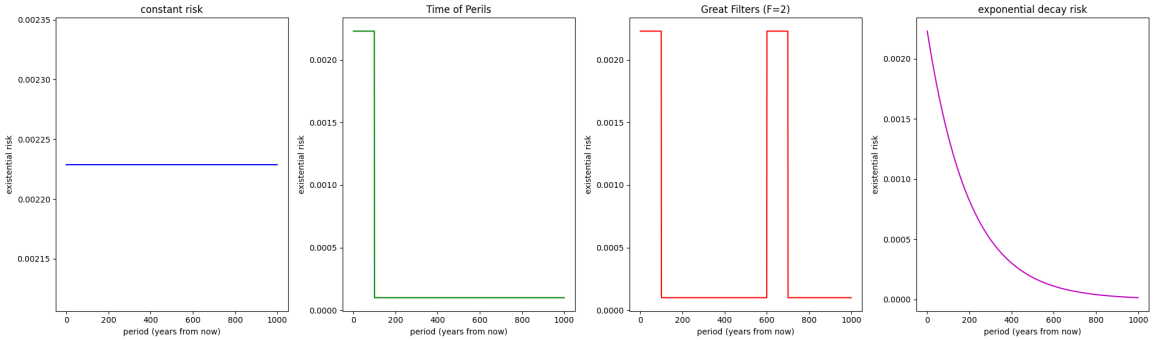


Figure 3: Risk Cases

Part II

Results

5 Convergence

As time goes to infinity, the expected value of existential risk mitigation could, in principle, be infinite. This would render comparing different estimates of $\mathbb{E}(M)$ redundant.[22] To investigate when this might happen, we turn our attention to convergence next.

We know that for any finite T, Equation 1 is bounded.[23] A key issue is whether the expected value of the world converges in an infinite universe. When $T \to \infty$, the series for the expected value of a world, $\mathbb{E}(w)$, as described in Equation 1, is given by the infinite sum

$$\mathbb{E}(w) = \sum_{t=1}^{\infty} \left[\left(\prod_{j=1}^{t} (1 - r_j) \right) v_t \right].$$

For this kind of series, we can use the Ratio Test to evaluate its convergence. The Ratio Test states that for a series $\sum_{n=1}^{\infty} a_n$, if there exists a limit

$$L = \lim_{n \to \infty} \left| \frac{a_{n+1}}{a_n} \right|,$$

then the series converges absolutely if $L < 1$, diverges if $L > 1$, and is inconclusive if $L = 1$.

To apply the Ratio Test to $\mathbb{E}(w)$, we look at consecutive terms of the series and their ratio.

$$\begin{aligned} L &= \lim_{t \to \infty} \left| \frac{\left(\prod_{j=1}^{t+1} (1 - r_j) \right) v_{t+1}}{\left(\prod_{j=1}^{t} (1 - r_j) \right) v_t} \right| \\ &= \lim_{t \to \infty} \left| \frac{\cancel{\left(\prod_{j=1}^{t} (1 - r_j) \right)} (1 - r_{t+1}) v_{t+1}}{\cancel{\left(\prod_{j=1}^{t} (1 - r_j) \right)} v_t} \right| \\ &= \lim_{t \to \infty} \left| \frac{v_{t+1}}{v_t} \right| (1 - r_{t+1}). \end{aligned}$$

Recall that $r_t \in (0, 1)$ for all i, so $(1 - r_t)$ also lies within $(0, 1)$ for all i. Thus, if r_t converges to a positive scalar, the exact risk level will not affect convergence. Instead, the convergence of the series $\mathbb{E}(w)$ critically depends on $\lim_{t \to \infty} \left| \frac{v_{t+1}}{v_t} \right|$. In particular, if this limit is less than or equal to 1, $\mathbb{E}(w)$ converges absolutely.[24] Therefore, we can write the following lemma stating a sufficient condition for the convergence of $\mathbb{E}(w)$.

[22] Tentatively, ordering infinite cardinalities could be a good option in those cases.
[23] For example by $T \cdot \max_{v_t} \{v_1, v_2, ...v_T\}$.
[24] More generally, the weakest condition we need to satisfy is that the limit is less than $1/(1 - r_t)$.

Lemma 1. *Suppose* $\lim_{t\to\infty} r_t$ *exists and is positive, then*

$$\mathbb{E}(w) \text{ converges if } V \equiv \lim_{t\to\infty} \left|\frac{v_{t+1}}{v_t}\right| \leq 1.$$

Proof. The Lemma follows by the reasoning above. Below is a more formal proof.

Suppose $V \equiv \lim_{t\to\infty} \left|\frac{v_{t+1}}{v_t}\right| \leq 1$. Then

$$\begin{aligned} L &= \lim_{n\to\infty} \left|\frac{a_{n+1}}{a_n}\right| \\ &= \lim_{t\to\infty} \left|\frac{v_{t+1}}{v_t}\right| (1 - r_{t+1}) \leq 1 \cdot \lim_{t\to\infty} (1 - r_{t+1}) \end{aligned}$$

And since $\lim_{t\to\infty} r_t$ exists and is positive, $\lim_{t\to\infty} (1 - r_{t+1}) < 1$. Thus $L \leq 1 \cdot \lim_{t\to\infty} (1 - r_{t+1})$; so $L < 1$. Hence, $\mathbb{E}(w)$ converges absolutely, and thus, it converges. □

We now investigate our main specific cases: constant, linear, quadratic, cubic, and polynomial sequences for v_t. We assume that $r_t \to r_\infty > 0$ throughout.

5.0.1 Constant $v_t = v_c$

For a constant sequence $v_t = v_c$, the series becomes:

$$\mathbb{E}(w) = \sum_{t=1}^{\infty} \left(\prod_{j=1}^{t} (1 - r_j) \right) v_c.$$

In particular, $V = v_c / v_c = 1$, so $\mathbb{E}(w)$ converges by Lemma 1.

5.0.2 Linear $v_t = v_c\, i$

For a linear sequence $v_t = v_c\, i$, we have:

$$\mathbb{E}(w) = \sum_{t=1}^{\infty} \left(\prod_{j=1}^{t} (1 - r_j) \right) v_c\, i.$$

Applying the Ratio Test, we find that $V = \lim_{t\to\infty} \frac{t+1}{i} = \lim_{t\to\infty} 1 + \frac{1}{i} = 1$, as the necessary condition for convergence requires.

5.0.3 Quadratic $v_t = v_c\ i^2$

In the quadratic case $v_t = v_c\ i^2$, we find:

$$V = \lim_{t\to\infty} \frac{(t+1)^2}{i^2} = \lim_{t\to\infty}(1 + \frac{2}{i} + \frac{1}{i^2}) = 1.$$

Therefore, the series converges.

5.0.4 Cubic $v_t = v_c\ i^3$

Similarly, for the cubic case $v_t = v_c\ i^3$,

$$V = \lim_{t\to\infty} \frac{(t+1)^3}{i^3} = 1,$$

applying Lemma 1 yields convergence.

5.0.5 Polynomial $v_t = v_c\ i^n$

The trend above continues: for any polynomial $v_t = v_c\ i^n$, $V = 1$ and the Ratio Test will yield $L < 1$, which yields absolute convergence.

5.0.6 Polynomial under Adjusted Sequences

When sequences are modified (see section 11) to replace i with $\alpha\ t + (1-\alpha)$, the Ratio Test yields the same outcomes. Consider the n-polynomial case $v_t = v_c\ (\alpha\ t + (1-\alpha))^n$ and evaluate V from Lemma 1 by repeatedly applying L'Hôpital Rule:

$$V = \lim_{t\to\infty} \frac{(\alpha\ (t+1) + (1-\alpha))^n}{(\alpha\ i + (1-\alpha))^n} = \lim_{t\to\infty} \frac{(\alpha\ (t+1) + (1-\alpha))^0}{(\alpha\ i + (1-\alpha))^0} = 1$$

Even with this modification, the series converges for constant, linear, quadratic, cubic, and n-polynomial value sequences.

5.0.7 Logistic Value

For logistic value we have $v_t = c/(1 + \frac{c-s}{s}e^{-\gamma t})$. Thus,

$$V = \lim_{t\to\infty}\left|\frac{v_{t+1}}{v_t}\right| = \lim_{t\to\infty}\frac{\frac{c}{1+\frac{c-s}{s}e^{-\gamma(t+1)}}}{\frac{c}{1+\frac{c-s}{s}e^{-\gamma t}}} = \lim_{t\to\infty}\frac{1+\frac{c-s}{s}e^{-\gamma t}}{1+\frac{c-s}{s}e^{-\gamma(t+1)}} = \frac{1+\frac{c-s}{s}\cdot 0}{1+\frac{c-s}{s}\cdot 0} = 1.$$

In the context of the various scenarios we've explored, we are now ready to present the following result:

Proposition 1. *The expected value of the world is finite if extinction risk converges to a any non-zero value (and if value growth is polynomial or logistic).*

Proof. The repeated application of Lemma 1 in the subsections above paired with the $r_t \to r_\infty > 0$ assumption, yields this proposition. □

Maintain the assumption that the risk tends to any nonzero value. As an immediate consequence of the above proposition, we have:

Corollary 1. *In an infinitely long universe, when* $\lim_{t\to\infty} r_t$ *exists and is positive, the value of extinction risk mitigation is finite.*

Proof.

$$\mathbb{E}(M) = \mathbb{E}(w') - \mathbb{E}(w)$$

and, by Proposition 1, both $\mathbb{E}(w')$ and $\mathbb{E}(w)$ converge. □

These results tell us that it is meaningful to talk about the long-term value of risk mitigation, even in the infinite universe case. Moreover, in all these scenarios, however great the value might be, it is simply not infinite. We estimate the exact size of this value in this Notebook.[25] Before that, we might wonder what the explicit closed formed solutions are for each scenario. We investigate those next.

6 Closed Form Solutions

When T is finite, the closed form equation for $\mathbb{E}(w)$ exists and directly follows from Equation 1.

Suppose that $T \to \infty$. We display a summary table with all the closed forms, with the derivations below it.

Let us first proceed without the adjustments

[25] It should be emphasise that the scope of 1 and 1 is the scenarios that this report considers, and not all the possible ways of modelling risk and value. For example, the proofs fail when the risk exponentially decays to zero, or when value grows exponentially without a cap.

6.1 Constant Risk

Constant risk is the simplest case, and it overlaps with previous work by Thorstad where he presents all but the cubic and logistic cases.

$$\sum_{t=1}^{\infty}\left[\left(\prod_{j=1}^{t}(1-r_j)\right)v_t\right]$$

with

$$v_t = i^n v_c \quad \text{and} \quad r(i) = r_c \in (0,1).$$

Thus we have

$$\mathbb{E}(w) = v_c \sum_{t=1}^{\infty}\left[(1-r_c)^t i^n\right]. \tag{11}$$

Let us start with the table:

Case	n	**Closed Form**
Constant	$n=0$	$v_c \cdot \frac{1-r_c}{r_c}$
Linear	$n=1$	$v_c \cdot \frac{1-r_c}{r_c^2}$
Quadratic	$n=2$	$v_c \cdot \frac{(1-r_c)(2-r_c)}{r_c^3}$
Cubic	$n=3$	$v_c \cdot \frac{(1-r_c)(1+4(1-r_c)+(1-r_c)^2)}{r_c^4}$
Polynomial	$n \in \mathbb{N}$	$v_c \cdot \left((1-r_c)\frac{\partial}{\partial(1-r_c)}\right)^n \frac{1-r_c}{r_c}$

Table 2: Closed-form solutions for polynomial value under constant risk

Derivations

1. For **constant**, $n=0$: Recall the expanded version of the series

$$\mathbb{E}(w) = v_c(1-r_c) + v_c(1-r_c)^2 + v_c(1-r_c)^3 + \dots$$

This is a geometric series with a first term $v_c(1-r_c)$ and a common ratio $(1-r_c) \in (0,1)$. We can apply the well known formula and obtain

$$\mathbb{E}(w) = v_c \cdot \frac{(1-r_c)}{1-(1-r_c)} = v_c \cdot \frac{(1-r_c)}{r_c}.$$

For $n > 0$ let us address the problem using the polylogarithm function (a general approach that works for $n \in \mathbb{N}$). The polylogarithm is defined as:

$$\text{Li}_s(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^s}$$

where s is the order of the polylogarithm and z is the complex argument where $|z| < 1$.

Let's consider the series in question

$$\frac{\mathbb{E}(w)}{v_c} = \sum_{t=1}^{\infty} i^n (1-r_c)^t,$$

which equals the expansion of the polylogarithm when setting $z = 1 - r_c, k = i$ and $s = -n$. To proceed, we use the well known expressions for the polylogarithm for the particular values $n = 1, 2, 3$ when $z = 1 - r_c$.

2. For **linear**, $n = 1$, from the known expression:

$$\text{Li}_{-1}(z) = \frac{z}{(1-z)^2} = \frac{1-r_c}{r_c^2}$$

when we substitute our value for z. So, the series becomes:

$$\mathbb{E}(w) = v_c \cdot \frac{1-r_c}{r_c^2}.$$

3. For **quadratic**, $n = 2$, we use

$$\text{Li}_{-2}(z) = \frac{z(1+z)}{(1-z)^3} = \frac{(1-r_c)(2-r_c)}{r_c^3}$$

substituting our value for z. The series is now

$$\mathbb{E}(w) = v_c \cdot \frac{(1-r_c)(2-r_c)}{r_c^3}.$$

4. For **cubic**, $n = 3$, use

$$\text{Li}_{-3}(z) = \frac{z(1+4z+z^2)}{(1-z)^4} = \frac{(1-r_c)(1+4(1-r_c)+(1-r_c)^2)}{r_c^4}.$$

And the series becomes:

$$\mathbb{E}(w) = v_c \cdot \frac{(1-r_c)(1+4(1-r_c)+(1-r_c)^2)}{r_c^4}.$$

For a more detailed section on closed form solutions, see the Appendix 12. It contains: an extension for a polynomial of general degree n, closed form solutions for the ToP, GFH, upper bounds for exponential decay risk and logistic growth, and a discussion of the closed form solutions for $\mathbb{E}(M)$.

7 The Expected Value of Mitigating Risk Visualised

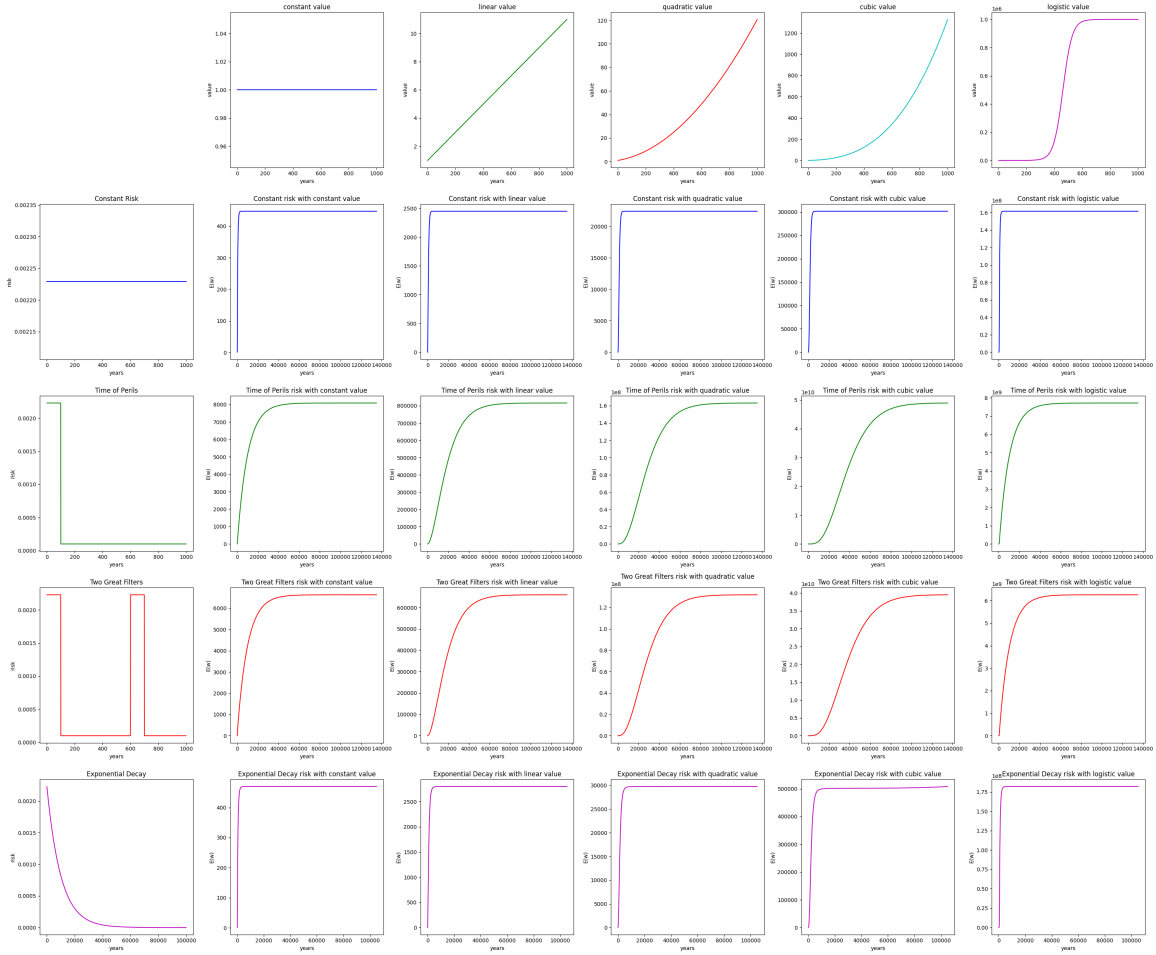


Figure 4: Grid: the value of the future

The first column indicates what value case we are on, the first row what risk case, and the middle plots display the cumulative $\mathbb{E}(w)$ as time passes for each risk and value

combination. Notice that in all cases, $\mathbb{E}(w)$ converges as $T \to \infty$. This is only indirectly related to section 5, which is about the convergence of $\mathbb{E}(M)$ and not the expected value of the future. For the middle plots, the horizontal axis displays the range from year zero (today), until year 140,000. For visibility, we display until year 100,000 for exponential decay instead. The vertical axis is different every time so that all graphs are clearly visible. For example, constant risk under linear value is in the thousands of v_c and Two Great Filters under logistic value is in billions of v_c, where v_c is always normalised to one. The default parameters for these simulations can be found and modified in the Notebook.

Next, we plot $\mathbb{E}(w)$, with and without performing M for all twenty scenarios in Figure 5. We do this for a range of persistence levels and, for entirely pedagogical reasons, we assume an extreme efficacy of $f = 50\%$ reduction in the risk from performing M.

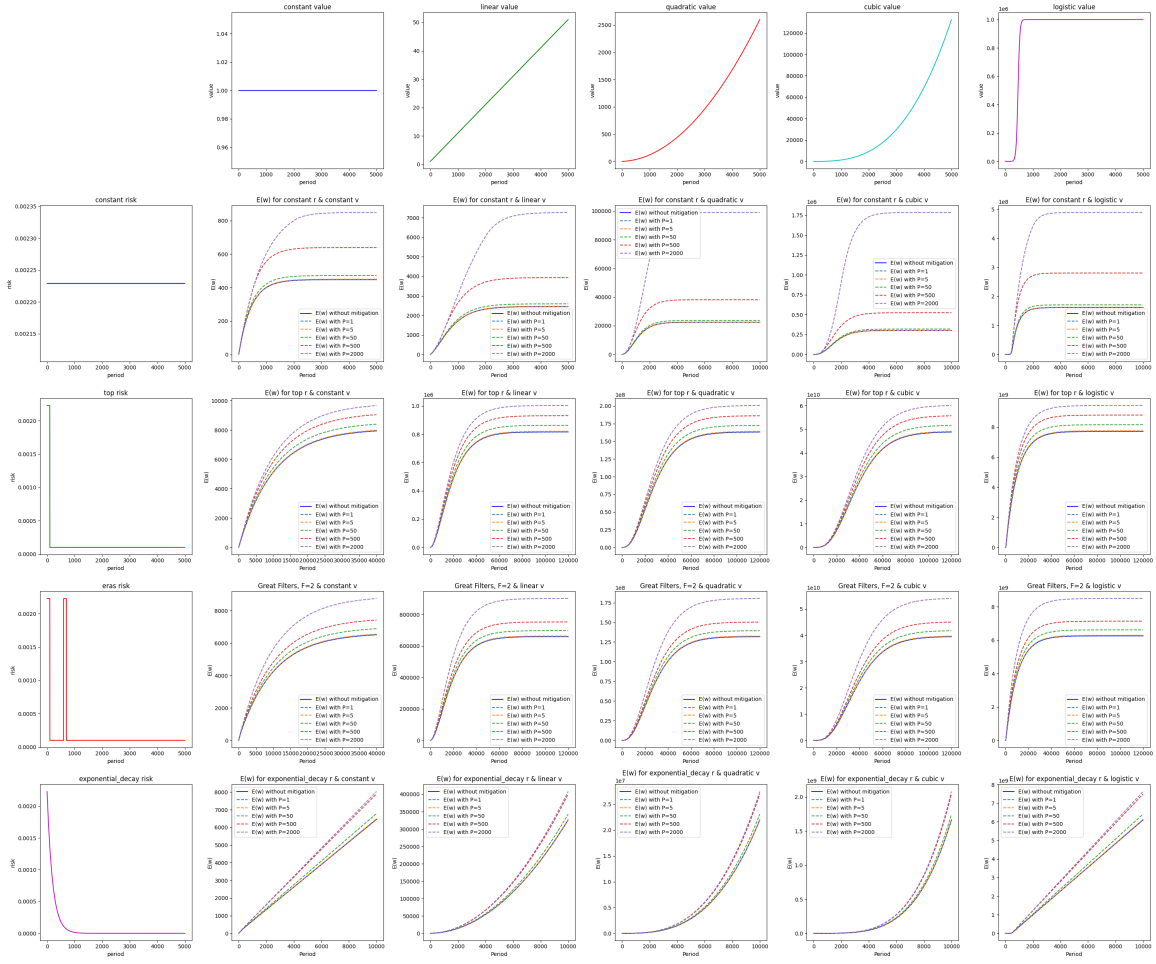


Figure 5: Grid: the value of mitigation when $f = 0.5$

In the grid above, to calculate $\mathbb{E}(M)$ for some specific case, we first take the dotted curve that tells us the expected value of the world after performing the action, all under a particular scenario and at certain persistence. Then, we subtract the baseline $\mathbb{E}(w)$

without mitigation, i.e. we subtract the solid blue curve from any one dotted curve.

When discussing the value and eventually the cost effectiveness of risk mitigation, a useful and more realistic efficacy f is one basis point: $f = 0.0001$.Table 3 below shows $\mathbb{E}(M)$ for all the scenarios of interest.

Model	Constant	Linear	Quadratic	Cubic	Logistic
			$P = 1$		
Constant	9.99×10^{-5}	5.48×10^{-4}	5.01×10^{-3}	6.74×10^{-2}	3.61×10^{1}
Time of Perils	1.77×10^{-3}	1.82×10^{-1}	3.64×10^{1}	1.09×10^{4}	1.72×10^{3}
2 Great Filters	1.46×10^{-3}	1.47×10^{-1}	2.94×10^{1}	8.82×10^{3}	1.40×10^{3}
Exponential Decay	1.05×10^{-4}	6.26×10^{-4}	6.64×10^{-3}	1.12×10^{-1}	4.06×10^{1}
			$P = 5$		
Constant	4.98×10^{-4}	2.74×10^{-3}	2.50×10^{-2}	3.37×10^{-1}	1.80×10^{2}
Time of Perils	8.87×10^{-3}	9.11×10^{-1}	1.82×10^{2}	5.46×10^{4}	8.62×10^{3}
2 Great Filters	7.28×10^{-3}	7.37×10^{-1}	1.47×10^{2}	4.41×10^{4}	6.99×10^{3}
Exponential Decay	5.22×10^{-4}	3.12×10^{-3}	3.32×10^{-2}	5.58×10^{-1}	2.03×10^{2}
			$P = 50$		
Constant	4.74×10^{-3}	2.71×10^{-2}	2.50×10^{-1}	3.37	1.80×10^{3}
Time of Perils	8.84×10^{-2}	9.11	1.82×10^{3}	5.46×10^{5}	8.62×10^{4}
2 Great Filters	7.25×10^{-2}	7.37	1.47×10^{3}	4.41×10^{5}	6.99×10^{4}
Exponential Decay	4.97×10^{-3}	3.09×10^{-2}	3.31×10^{-1}	5.57	2.03×10^{3}
			$P = 500$		
Constant	3.02×10^{-2}	2.27×10^{-1}	2.37	3.33×10^{1}	1.78×10^{4}
Time of Perils	2.07×10^{-1}	2.15×10^{1}	4.30×10^{3}	1.29×10^{6}	2.03×10^{5}
2 Great Filters	1.70×10^{-1}	1.74×10^{1}	3.47×10^{3}	1.04×10^{6}	1.65×10^{5}
Exponential Decay	3.19×10^{-2}	2.59×10^{-1}	3.11	5.40×10^{1}	1.96×10^{4}
			$P = 2000$		
Constant	4.43×10^{-2}	4.31×10^{-1}	5.76	9.83×10^{1}	3.20×10^{4}
Time of Perils	3.12×10^{-1}	3.36×10^{1}	6.74×10^{3}	2.02×10^{6}	3.10×10^{5}
2 Great Filters	3.82×10^{-1}	4.12×10^{1}	8.26×10^{3}	2.47×10^{6}	3.82×10^{5}
Exponential Decay	4.85×10^{-2}	5.20×10^{-1}	8.02	1.67×10^{2}	3.62×10^{4}

Table 3: $\mathbb{E}(M)$ for all scenarios when $f = 0.0001$

Though we show it above, we are suspicious of long persistence, both because effects are blunted by political or technological changes and because, given enough time, some actor is likely to perform an action that achieves similar effects.[26]

Given the difference in orders of magnitude, it can be difficult to directly compare the figures in this table. To facilitate this, we display Figure 6: a visual representation of the estimated expected value of reducing existential risk by 0.01%.[27] The image is to scale and one cubic unit is the size of the world under constant risk and constant value, the top-left scenario. A persistence of 5 years is assumed.

[26]On the latter point, to calculate the actual difference that our efforts makes to the effects of persistence will require future work. For example, imagine you do an action, M, at $t = 1$ that mitigates risk for the next 10 years. If you hadn't done M, someone else would have taken that same action at $t = 5$. How should we measure the persistence and value of M in this case? The treatment of 'contingency' here can help guide our thoughts.

[27]Because of computational limits, the expected value calculation assumes a cap of 120 thousand years. This is more than long enough in most scenarios, where a T this large achieves the same behaviour as $T \to \infty$, but nuances arise in the exponential decay case, see the notebook for a thorough discussion of those.

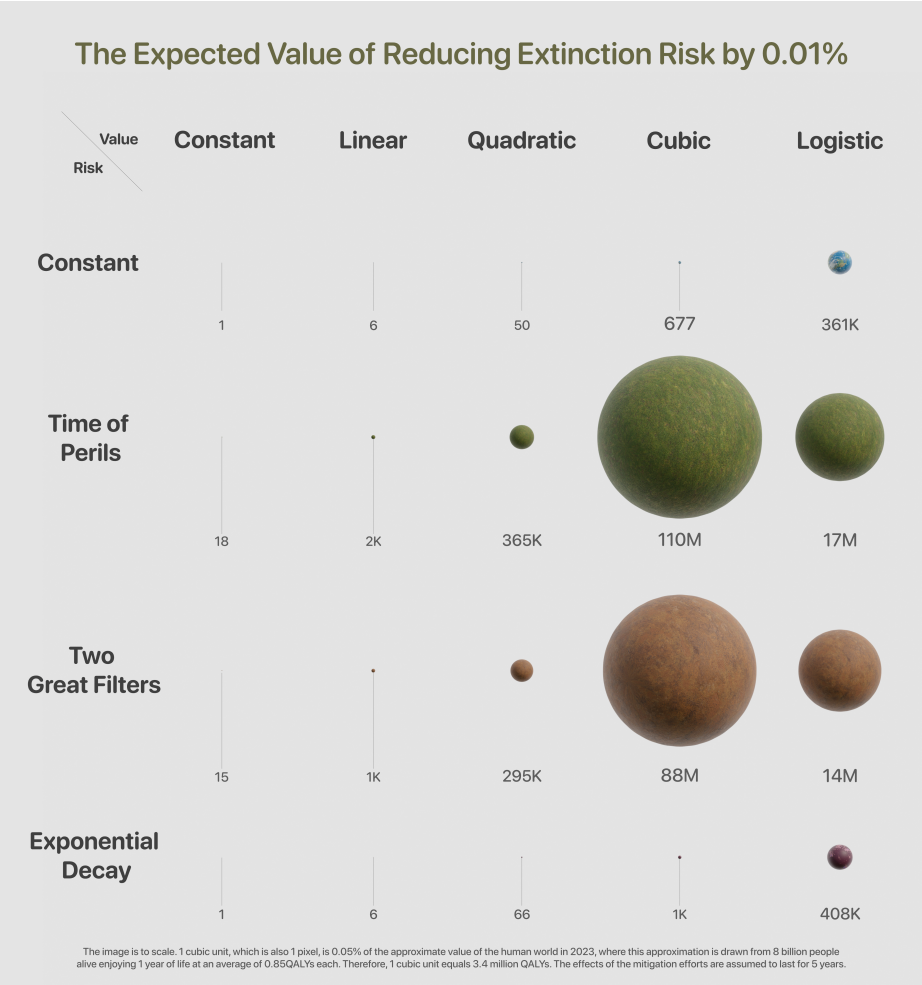


Figure 6: $\mathbb{E}(M)$ when $f = 0.0001$

For an extended discussion of these results see the full report. Here are some key takeaways:

- How many orders of magnitude $\mathbb{E}(M)$ is under Time of Perils crucially depends on assumptions about value growth (it is 11 million times bigger under cubic value compared to constant).
- For constant value, as we vary the assumed risk and persistence, $\mathbb{E}(M)$ stays within one order of magnitude above or below the median value in Table 3. For linear and quadratic it's within two orders of magnitude.
- Adding another filter keeps $\mathbb{E}(M)$ in the same order of magnitude, and only reduces it by about 25%, under the default parameters in the Notebook.
- Given a fixed persistence, there's still extreme variability: the minimum $\mathbb{E}(M)$ is roughly 8 orders of magnitude smaller than the maximum.

- This extreme difference can be put succinctly: suppose that the units were meters travelled as you walk away from London Bridge. The smallest value implies you'd walk 17cm, about the length of a pencil. Whereas the largest means that you'd walk from London to Sydney.

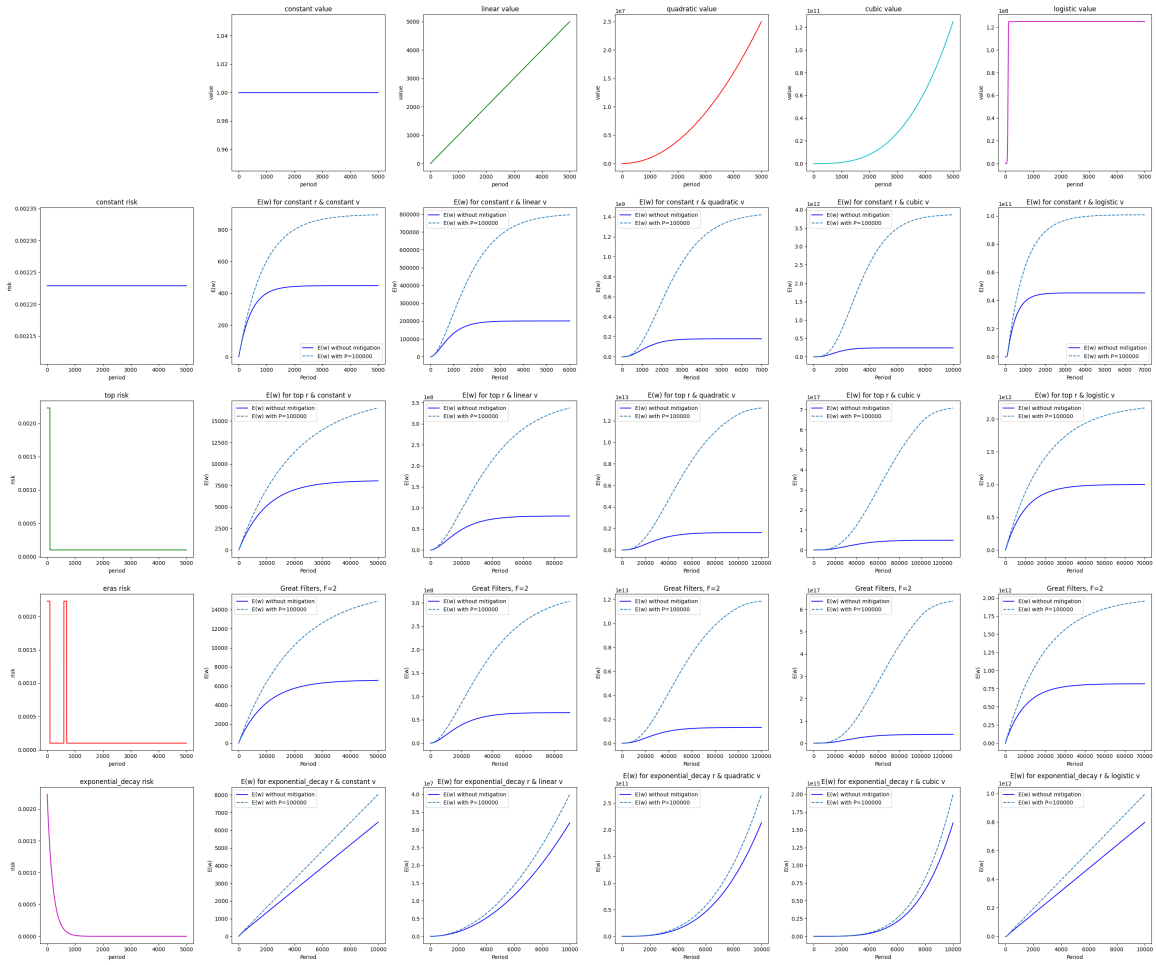


Figure 7: Grid: the value of mitigation for large P

The Role of Persistence

Two remarks seem worth making. First, that persistence plays a key role in the value of risk mitigation. For example, in Figure 8 below, depending on persistence $\mathbb{E}(M)$ can increase by up to 30 times. Second, we suggest an empirical hypothesis that persistence is unlikely to be higher than 50 years. The reasoning here is that there might be interventions that reduce risk a lot for not very long or not very much but for a long time. But actions that drastically reduce risk and do so for a long time are rare. Jointly these two remarks entail that the value of risk mitigation is between one ten-thousandth of a v_c (under constant risk and value) and two billion v_c (under cubic

and time of perils assuming f is one basis point), a considerable range.[28]

To illustrate the role of persistence consider the following picture, which plots $\mathbb{E}(w)$ versus persistence in the constant risk and value case for $f = 0.0001$.

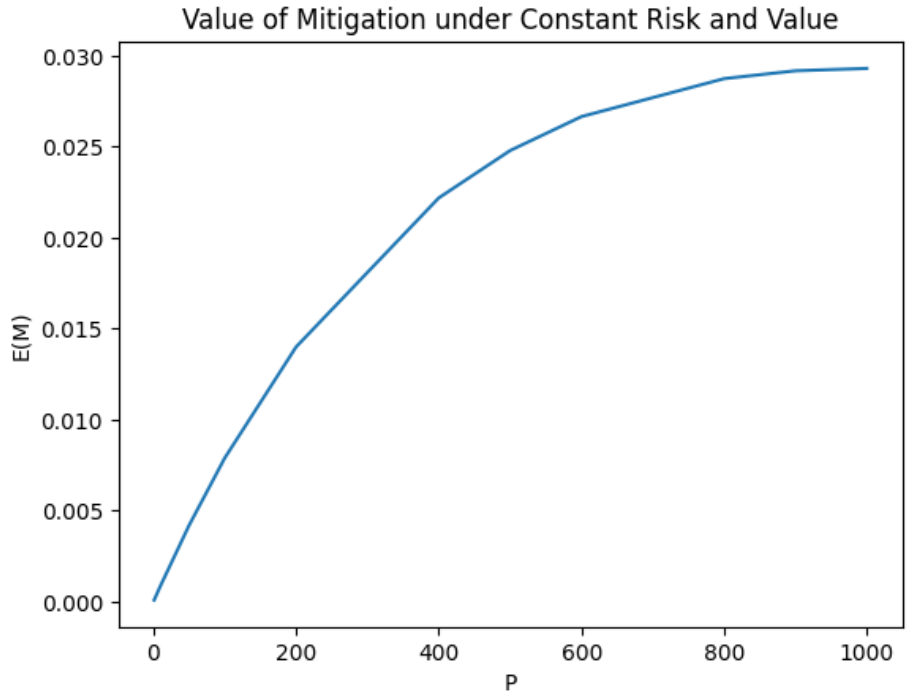


Figure 8: Plot of $\mathbb{E}(w)$ versus persistence for constant risk and value.

Increasing persistence is important but it exhibits decreasing marginal returns in the concave fashion illustrated above.

This result matches our intuitions. Because of its cumulative nature, the probability of avoiding extinction in the near-term is much higher than avoiding it long-term. That means that the value contributions to $\mathbb{E}(w)$, which also impact $\mathbb{E}(M)$, are much higher in the short term than in the long term, when they are heavily discounted by the probability of them taking place. So the marginal gains from increasing persistence are much higher in the short term than in the long term. In other words, for example, adding 1 year of persistence to a mitigation action whose effects last 1 year is much more valuable than adding 1 year of persistence to a mitigation action whose effects last 100 years. A general lesson follows: performing actions that have larger persistence is key, but increasing persistence is particularly valuable for low persistence values.

[28] Recall previous footnote defining v_c.

8 Concluding Remarks

This report is restricted in its scope and has a number of limitations. If there is enough value and interest in this type of work, our follow-up research could include:

- a friendlier online platform with sliders and buttons to select and tweak the scenarios users want to visualise
- explicit closed form expressions for comparative statics, formulae that describe the impact of shifting key parameters on $\mathbb{E}(M)$
- explicit uncertainty analyses with Monte Carlo simulations where we graphically observe the importance of key parameters and different upper and lower bounds of $\mathbb{E}(M)$ according to a range of scenarios
- more sophisticated treatments of persistence
- discussions about option value and its role in thinking about existential risk mitigation
- modelling efforts that improve value trajectory and could be competitive with extinction risk reduction
- including partial catastrophes
- formally exploring other events conceptually included in existential risk but not extinction risk
- including population growth as a parameter that directly affects values
- new scenarios, including explicit treatment of population growth and other non-human sentience
- investigating value trajectories that feature negative value

With these limitations in mind, some **points of caution about practical upshots** include:

- Depending on the parameters of exponential decay, and the time horizon, convergence under exponential decay risk can be misleading, check the Jupyter Notebook for full details.[29]
- While the results here might help us arrive at better-informed expected value judgements, this report is not meant to settle questions about how to form an overarching view on the overall value of extinction risk mitigation. A lot more work is needed for that, for instance, our views on risk aversion could have a sizeable impact on this.
- Be careful with using the reports' results to perform back of the envelope calculations with new parameters in mind, and update your views by roughly deducting or adding some orders of magnitude. When possible, rerun the code instead.[30]

[29] In particular, Figure 6's exponential decay values were approximated using the first 100,000 years.
[30] I'm happy to help with this.

- More broadly, while a more complex model like this one can certainly model things that were previously left out, we have so little data to fit it to that we should be especially cautious about over-updating from specific quantitative conclusions.

This report extended the model developed by Ord, Thorstad and Adamczewski. By enriching the base model, we were able to perform sensitivity analyses, observe convergence and can now better evaluate when extinction risk mitigation could, in expectation, be overwhelmingly valuable, and when it is comparable to or of lesser value than the alternatives. Crucially, we show that the value of extinction risk work varies considerably with different assumptions about the relevant risk and value scenarios. Insofar as we don't have much confidence in any one scenario, we should form views that reflect this uncertainty and we shouldn't have much confidence in any particular estimate of the value of risk mitigation efforts.

9 References

References

1977QJRAS..18....3I Page 3 (2023). URL: https://adsabs.harvard.edu/full/1977QJRAS..18....3I.

Adamczewski, Tom (2023). *The expected value of the long-term future, and existential risk*. URL: https://bayes.net/ltf-paper/.

Adams, Fred C. and Gregory Laughlin (Apr. 1, 1997). "A dying universe: the long-term fate and evolutionof astrophysical objects". In: *Reviews of Modern Physics* 69.2. Publisher: American Physical Society, pp. 337–372. DOI: 10.1103/RevModPhys.69.337. URL: https://link.aps.org/doi/10.1103/RevModPhys.69.337.

Aschenbrenner, Leopold (n.d.). "Existential Risk and Growth". In: ().

Bostrom, Nick (Nov. 2003). "Astronomical Waste: The Opportunity Cost of Delayed Technological Development". In: *Utilitas* 15.3. Publisher: Cambridge University Press, pp. 308–314. ISSN: 1741-6183, 0953-8208. DOI: 10.1017/S0953820800004076. URL: https://www.cambridge.org/core/journals/utilitas/article/abs/astronomical-waste-the-opportunity-cost-of-delayed-technological-development/2969D64410332BD099F36BAFC5B2ADE5

Broome, John (1994). "Discounting the Future". In: *Philosophy & Public Affairs* 23.2, pp. 128–156. ISSN: 1088-4963. DOI: 10.1111/j.1088-4963.1994.tb00008.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1088-4963.1994.tb00008.x.

Davies, Paul (Jan. 9, 1997). *The Last Three Minutes: Conjectures About The Ultimate Fate Of The Universe*. Basic Books. ISBN: 978-0-465-03851-0.

Dyson, Freeman J. (July 1, 1979). "Time without end: Physics and biology in an open universe". In: *Reviews of Modern Physics* 51.3, pp. 447–460. ISSN: 0034-6861. DOI: 10.1103/RevModPhys.51.447. URL: https://link.aps.org/doi/10.1103/RevModPhys.51.447.

elifland (2023). "Prioritizing x-risks may require caring about future people". In: (). URL: https://forum.effectivealtruism.org/posts/rvvwCcixmEep4RSjg/prioritizing-x-risks-may-require-caring-about-future-people.

Ellis, George Francis Rayner (2002). *The Far-future Universe: Eschatology from a Cosmic Perspective*. Google-Books-ID: WN8L3pwfpW8C. Templeton Foundation Press. 412 pp. ISBN: 978-1-890151-90-4.

Jones, Charles I (n.d.). "The A.I. Dilemma: Growth versus Existential Risk". In: ().
Krauss, Lawrence M. and Glenn D. Starkman (1999). "The Fate of Life in the Universe". In: *Scientific American* 281.5. Publisher: Scientific American, a division of Nature America, Inc., pp. 58–65. ISSN: 0036-8733. URL: https://www.jstor.org/stable/26058484.
Lewis, Gregory (2023). "The person-affecting value of existential risk reduction". In: (). URL: https://forum.effectivealtruism.org/posts/dfiKak8ZPa46N7Np6/the-person-affecting-value-of-existential-risk-reduction.
Linch (2023a). ".01% Fund - Ideation and Proposal". In: (). URL: https://forum.effectivealtruism.org/posts/pGaesfn4R9xWK4Rmk/01-fund-ideation-and-proposal.
— (2023b). "How many EA 2021 $s would you trade off against a 0.01% chance of existential catastrophe?" In: (). URL: https://forum.effectivealtruism.org/posts/cKPkimztzKoCkZ75r/how-many-ea-2021-usds-would-you-trade-off-against-a-0-01.
MacAskill, William, Teruji Thomas, and Aron Vallinder (n.d.). "The Significance, Persistence, Contingency Framework". In: ().
Matheny, Jason G. (2007). "Reducing the Risk of Human Extinction". In: *Risk Analysis* 27.5. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1539-6924.2007.00960.x, pp. 1335–1344. ISSN: 1539-6924. DOI: 10.1111/j.1539-6924.2007.00960.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1539-6924.2007.00960.x.
Millett, Piers and Andrew Snyder-Beattie (Aug. 1, 2017). "Existential Risk and Cost-Effective Biosecurity". In: *Health Security* 15.4, pp. 373–383. ISSN: 2326-5094. DOI: 10.1089/hs.2017.0028. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5576214/.
Ord, Toby (2020). *The Precipice: Existential Risk and the Future of Humanity*. London: Bloomsbury.
Physics, Mr Turnbull's (Mar. 28, 2021). *The Fate of the Universe*. URL: https://www.youtube.com/watch?v=N5qafppFntU.
Tarsney, Christian (May 31, 2023a). "The epistemic challenge to longtermism". In: *Synthese* 201.6, p. 195. ISSN: 1573-0964. DOI: 10.1007/s11229-023-04153-y. URL: https://link.springer.com/10.1007/s11229-023-04153-y.
— (May 31, 2023b). "The epistemic challenge to longtermism". In: *Synthese* 201.6, p. 195. ISSN: 1573-0964. DOI: 10.1007/s11229-023-04153-y. URL: https://link.springer.com/10.1007/s11229-023-04153-y.
The Great Filter (2023). URL: http://mason.gmu.edu/~rhanson/greatfilter.html.
Thorstad, David (Aug. 22, 2023). "High Risk, Low Reward: A Challenge to the Astronomical Value of Existential Risk Mitigation". In: *Philosophy & Public Affairs*, papa.12248. ISSN: 0048-3915, 1088-4963. DOI: 10.1111/papa.12248. URL: https://onlinelibrary.wiley.com/doi/10.1111/papa.12248.
Toby_Ord (2023). "Shaping Humanity's Longterm Trajectory". In: (). URL: https://forum.effectivealtruism.org/posts/Doa69pezbZBqrcucs/shaping-humanity-s-longterm-trajectory.
Why despite global progress, humanity is probably facing its most dangerous time ever (2023). 80,000 Hours. URL: https://80000hours.org/articles/existential-risks/.

10 Acknowledgements

The post was written by Arvo Muñoz Morán. Thank you to the members of the Worldview Investigations Team – David Bernard, Hayley Clatterbuck, Bob Fischer, Laura Duffy and Derek Shiller – Marcus Davis, Toby Ord, Elliott Thornley, Tom Houlden, Loren Fryxell, Lucy Hampton, Adam Binks, Jacob Peacock and Daniel Carey for helpful comments and discussions. The post is a project of – Rethink Priorities, a global priority think-and-do tank, aiming to do good at scale. We research and implement pressing opportunities to make the world better. We act upon these opportunities by developing and implementing strategies, projects, and solutions to key issues. We do this work in close partnership with foundations and impact-focused non-profits or other entities. If you're interested in Rethink Priorities' work, please consider subscribing to our newsletter. You can explore our completed public work here.

11 Century to Year Adjustments

Consider the problem of adapting a sequence originally defined in terms of centuries (like in previous versions of the models) to a sequence defined in terms of years, ensuring that the behaviour of the sequence remains consistent.[31]

Linear Case

Let us begin with the simplest scenario of a linear sequence, $v(i) = iv_c$, where $i \in \mathbb{Z}^+$ was the index in centuries. To represent this sequence in years, denoted $y \in \mathbb{Z}^+$, we introduce the parameter $\alpha \in (0, 1]$ and the transformation

$$v(t) = (\alpha t + (1 - \alpha))\, v_c.$$

For the unit conversion from centuries to years, α is chosen as $1/100$. This yields $v(y) = \left(\frac{1}{100}y + 0.99\right) v_c$. Here is a table to illustrate the transformation.

Table 4: Transformation of Linear Sequence from Centuries to Years

Year y	$v(y)$	Corresponding i in Centuries
1	v_c	1
51	$1.5v_c$	
101	$2v_c$	2
201	$3v_c$	3
301	$4v_c$	4
801	$9v_c$	9
901	$10v_c$	10
1,001	$11v_c$	11
100,001	$1,001v_c$	1,001

On the 51st year, the value of the world is 50% larger than it used to. On the 101st year, with the beginning of century number 2, the value has now doubled, reaching $2v_c$. We should note the peculiarity that after 900 years, we are technically on the 10th century, and so the value has increased tenfold. After 100,000 years, Table 4 shows how the value is 1001.

Quadratic Case

Next, consider a quadratic sequence defined as $v(i) = i^2 v_c$, with i still representing the index in centuries. Utilising the same α-based transformation, the sequence in terms of years becomes $v(y) = \left(\frac{1}{100}y + 0.99\right)^2 v_c$. For pedagogical purposes, we include Table 5, which is entirely analogous to Table 4.

Cubic Case

[31] To see why this is necessary, consider the cubic case. While it is already quite optimistic to suppose that the next century will see 8 times the value of the present and the one after that 27 times, it is too unrealistic to assume the same phenomena over the next three years. With that in mind, we want the tools to match the growth in centuries.

Table 5: Transformation of Quadratic Sequence from Centuries to Years

Year y	$v(y)$	Corresponding i in Centuries
1	v_c	1
51	$1.5^2 v_c$	1
101	$2^2 v_c$	2
801	$9^2 v_c$	9
901	$10^2 v_c$	10
1,001	$11^2 v_c$	11
100,001	$1,001^2 v_c$	1,001

For a cubic sequence $v(i) = i^3 v_c$, the representative element of the sequence in years is $v(y) = \left(\frac{1}{100}y + 0.99\right)^3 v_c$.

Generalisation to Polynomial Growth of Order n

The approach outlined above generalises naturally to sequences exhibiting polynomial growth of order n. For a sequence in centuries represented by $v(i) = i^n v_c$, the corresponding sequence in years becomes $v(y) = \left(\frac{1}{100}y + 0.99\right)^n v_c$.

We introduced α as a parameter to handle unit conversion, and it plays the pivotal role in adjusting the sequence across different timescales with the transformation $v(t) = (\alpha t + (1-\alpha))^n v_c$. For converting from centuries to years, $\alpha = 1/100$, while for decades to years, $\alpha = 1/10$.

This methodology provides a consistent way to represent value growth sequences across varying timescales, moreover, α can be adjusted as a the rate of growth parameter to represent any polynomial growth trajectory desired. For example, one might think that cubic with a lower alpha $\alpha = 0.00001$ would be the most realistic trajectory.

12 Other Closed Forms

12.1 Constant risk with polynomial value of order n

For a general n, we have

$$\text{Li}_{-n}(z) = \left(z\frac{\partial}{\partial z}\right)^n \frac{z}{1-z} = \sum_{k=0}^{n} k! S(n+1, k+1) \left(\frac{z}{1-z}\right)^{k+1},$$

where $n \in \mathbb{N}$ and $S(n,k)$ are the Stirling numbers of the second kind.

The Stirling numbers of the second kind, represented as $S(n,k)$ or $\left\{ \begin{array}{c} n \\ k \end{array} \right\}$, are significant combinatorial numbers that count the ways to partition a set of n labelled objects

into k nonempty unlabelled subsets. Their formula is given by:

$$\left\{ \begin{array}{c} n \\ k \end{array} \right\} = \frac{1}{k!} \sum_{t=0}^{k} (-1)^{k-i} \binom{k}{i} i^n = \sum_{t=0}^{k} \frac{(-1)^{k-i} i^n}{(k-i)!i!}.$$

Setting $z = 1 - r_c, k = 1$:

$$\text{Li}_{-n}(1-r_c) = \left((1-r_c) \frac{\partial}{\partial (1-r_c)} \right)^n \frac{1-r_c}{r_c} = \sum_{t=0}^{n} i! S(n+1, t+1) \left(\frac{1-r_c}{r_c} \right)^{t+1},$$

and for the Stirling numbers expression

$$\left\{ \begin{array}{c} n \\ i \end{array} \right\} = \frac{1}{i!} \sum_{j=0}^{t} (-1)^{i-j} \binom{i}{j} j^n = \sum_{j=0}^{t} \frac{(-1)^{i-j} j^n}{(i-j)!j!}.$$

12.2 Constant Risk closed forms given yearly adjustments

Keep the constant risk assumption but consider instead, a setting where $v_t = (\alpha i + (1-\alpha))^n v_c$, to reflect our value growth adjustments in section 11.

Tentatively[32], the closed form solutions are as follows.

12.2.1 Constant

The adjustment has no impact, it is the same as before.

12.2.2 Linear

Given the series:

$$\mathbb{E}(w) = \sum_{t=1}^{\infty} \left(\prod_{j=1}^{t} (1-r_j) \right) v_t$$

For $n = 1$:

$$v_t = (\alpha i + (1-\alpha)) v_c$$

We'll use the geometric series sum formula to find the closed form for this series.

[32] i.e. liable to typos.

To clarify, for our problem $r_j = r_c$ for all j. For each term, considering only the risk factors:

$$\prod_{j=1}^{t}(1 - r_j) = (1 - r_c)^t$$

Given the above, we can express our series with the new value function as:

$$\mathbb{E}(w) = \sum_{t=1}^{\infty}(1 - r_c)^t (\alpha i + (1 - \alpha)) v_c$$

Separating the summation:

$$\mathbb{E}(w) = v_c \sum_{t=1}^{\infty} \alpha i (1 - r_c)^t + v_c(1 - \alpha) \sum_{t=1}^{\infty}(1 - r_c)^t$$

Let's solve the two summations:

1. For the first sum, consider $\sum_{t=1}^{\infty} ix^t$, this is a known series and its sum is $\frac{x}{(1-x)^2}$, when $|x| < 1$. For our case, $x = (1 - r_c)$.

2. The second sum is simply the geometric series and its sum is $\frac{x}{1-x}$, when $|x| < 1$. For our case, $x = (1 - r_c)$.

Substituting these in, we get:

$$\mathbb{E}(w) = v_c \left[\alpha \frac{(1 - r_c)}{{r_c}^2} + (1 - \alpha) \frac{(1 - r_c)}{r_c} \right].$$

Simplifying

$$\mathbb{E}(w) = v_c \frac{(1 - r_c)}{r_c} \left[\alpha \frac{1 - r_c}{r_c} + 1 \right].$$

12.2.3 Quadratic

Alright! Let's tackle the case $n = 2$.

Given:

$$v_t = (\alpha i + (1 - \alpha))^2 v_c$$

For $n = 2$:

$$v_t = \left(\alpha^2 i^2 + 2\alpha(1 - \alpha) i + (1 - \alpha)^2 \right) v_c$$

Now, as before, our series expression becomes:

$$\mathbb{E}(w) = \sum_{t=1}^{\infty}(1 - r_c)^t \left(\alpha^2 i^2 + 2\alpha(1 - \alpha) i + (1 - \alpha)^2 \right) v_c$$

This can be separated into three summations:

1. $\alpha^2 v_c \sum_{t=1}^{\infty} i^2 (1-r_c)^t$

2. $2\alpha(1-\alpha) v_c \sum_{t=1}^{\infty} i (1-r_c)^t$

3. $(1-\alpha)^2 v_c \sum_{t=1}^{\infty} (1-r_c)^t$

The second and third sums we already tackled in the previous $n = 1$ case. The new sum to evaluate here is the first one:

For $\sum_{t=1}^{\infty} i^2 x^t$, using differentiation properties of power series, or polylogarithms like before, its sum can be obtained as:

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

provided $|x| < 1$. For our problem, $x = (1 - r_c)$.

Substituting into our summation:

1. $\alpha^2 v_c \sum_{t=1}^{\infty} i^2 (1-r_c)^t = \alpha^2 v_c \frac{(1-r_c)(2-r_c)}{r_c^3}$

2. $2\alpha(1-\alpha) v_c \sum_{t=1}^{\infty} i (1-r_c)^t = 2\alpha(1-\alpha) v_c \frac{(1-r_c)}{r_c^2}$

3. $(1-\alpha)^2 v_c \sum_{t=1}^{\infty} (1-r_c)^t = (1-\alpha)^2 v_c \frac{(1-r_c)}{r_c}$

Combining these, the closed form for the expected value $\mathbb{E}(w)$ when $n = 2$ is:

$$\mathbb{E}(w) = v_c \left[\alpha^2 \frac{(1-r_c)(2-r_c)}{r_c^3} + 2\alpha(1-\alpha) \frac{(1-r_c)}{r_c^2} + (1-\alpha)^2 \frac{(1-r_c)}{r_c} \right].$$

12.2.4 Cubic

Given:

$$v_t = (\alpha i + (1-\alpha))^3 v_c$$

For $n = 3$:

$$v_t = \left(\alpha^3 i^3 + 3\alpha^2 (1-\alpha) i^2 + 3\alpha (1-\alpha)^2 i + (1-\alpha)^3 \right) v_c$$

Now, our series expression becomes:

$$\mathbb{E}(w) = \sum_{t=1}^{\infty} (1-r_c)^t \left(\alpha^3 i^3 + 3\alpha^2 (1-\alpha) i^2 + 3\alpha (1-\alpha)^2 i + (1-\alpha)^3 \right) v_c$$

This can be separated into four summations:

1. $\alpha^3 v_c \sum_{t=1}^{\infty} i^3 (1-r_c)^t$

2. $3\alpha^2(1-\alpha) v_c \sum_{t=1}^{\infty} i^2 (1-r_c)^t$

3. $3\alpha(1-\alpha)^2 v_c \sum_{t=1}^{\infty} i (1-r_c)^t$

4. $(1-\alpha)^3 v_c \sum_{t=1}^{\infty} (1-r_c)^t$

We already tackled the second, third, and fourth sums in the previous $n = 1$ and $n = 2$ cases.

The new challenge is to evaluate the first sum. For the series $\sum_{t=1}^{\infty} i^3 x^t$, we can use the following identity:

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

provided $|x| < 1$. For our problem, $x = (1-r_c)$.

Substituting into our summation:

1. $\alpha^3 v_c \sum_{t=1}^{\infty} i^3 (1-r_c)^t = \alpha^3 v_c \frac{(1-r_c)(1+4(1-r_c)+(1-r_c)^2)}{r_c^4}$

2. $3\alpha^2(1-\alpha) v_c \sum_{t=1}^{\infty} i^2 (1-r_c)^t = 3\alpha^2(1-\alpha) v_c \frac{(1-r_c)(2-r_c)}{r_c^3}$

3. $3\alpha(1-\alpha)^2 v_c \sum_{t=1}^{\infty} i (1-r_c)^t = 3\alpha(1-\alpha)^2 v_c \frac{(1-r_c)}{r_c^2}$

4. $(1-\alpha)^3 v_c \sum_{t=1}^{\infty} (1-r_c)^t = (1-\alpha)^3 v_c \frac{(1-r_c)}{r_c}$

Combining these, the closed form for the expected value $\mathbb{E}[w]$ when $n = 3$ is:

$$\mathbb{E}[w] = v_c[\alpha^3 \frac{(1-r_c)(1+4(1-r_c)+(1-r_c)^2)}{r_c^4} + 3\alpha^2(1-\alpha)\frac{(1-r_c)(2-r_c)}{r_c^3} + 3\alpha(1-\alpha)^2\frac{(1-r_c)}{r_c^2} + (1-\alpha)^3\frac{(1-r_c)}{r_c}]$$