Existential Risk and Growth

Philip Trammell* and Leopold Aschenbrenner†

December 23, 2025

Technological development raises consumption but may pose existential risk. A growing literature studies this tradeoff in static settings where stagnation is perfectly safe. But if any risky technology already exists, technological development can also lower risk indirectly in two ways: by speeding (1) technological solutions and/or (2) a "Kuznets curve" in which wealth increases a planner's willingness to pay for safety. The risk-minimizing technology growth rate, in light of these dynamics, is typically positive and may easily be high. Below this rate, technological development poses no tradeoff between consumption and cumulative risk.

*Digital Economy Lab, Stanford University. Contact: trammell@stanford.edu. We thank Danny Bressler, Lennart Stern, and Michael Wiebe for suggesting the idea; Ben Snodin, Luis Mota, Tyler Cowen, Rick van der Ploeg, Pete Klenow, Chad Jones, Toby Ord, Seth Benzell, Ioannis Bournakis, Bård Harstad, Andrew Koh, and attendees of several workshops and work-in-progress seminars at the Global Priorities Institute for helpful comments; and Alex Holness-Tofts and Arvo Muñoz for assistance on an earlier draft. This paper was written with support from Open Philanthropy, the Future of Humanity Institute (University of Oxford), and the Centre for Effective Altruism.

†Situational Awareness LP.

1 Introduction

Technology increases consumption. It may also pose *existential risk*, or “x-risk”: the risk of human extinction or, equivalently for decision purposes, an equally complete and permanent welfare loss.[1] Advanced biotechnology (Millett and Snyder-Beattie, 2017), nuclear weapons (Geist et al., 2024),[2] and emissions-intensive industrial production (Steffen et al., 2018) have been argued to pose such risks, and x-risk from AI is now also a widespread concern (Center for AI Safety, 2023; Jones, 2024, 2025).

This suggests a tradeoff between x-risk and consumption growth. Bostrom (2003) argues that if we do not discount the welfare of future generations, x-risk looms especially large: the benefits of saving the world could last almost indefinitely, whereas speeding technological development only yields significant benefits in the “short term”, by pulling forward the time when the pool of useful technologies is exhausted. Ord (2024) offers a helpful exposition of this and related points. Baranzini and Bourguinion (1995) consider the “growth vs. risk” tradeoff within a more conventional economic framework, focusing on the conditions under which the optimal policy is also safest. Nordhaus (2011), Méjean et al. (2020), and Jones (2016, 2024, 2025) more generally study the amount of consumption worth sacrificing for existential safety.[3]

To our knowledge, every economic model to date of the impact of growth[4] on x-risk assumes that stagnation is perfectly safe. This condition is extreme and arguably unrealistic. Even if we developed no new technology, our ability to develop and deploy nuclear and biological weapons, and/or the possibility of triggering a runaway

[1]See e.g. Bostrom (2002), Posner (2004), Farquhar et al. (2017), and Ord (2020). We will refer to the event that humanity immediately goes extinct or suffers a similarly complete and permanent welfare loss as an “existential catastrophe”, or simply “catastrophe”. Our definition excludes gradual events such as slow AI takeover (Christiano, 2019; Kulveit et al., 2025). We will refer to “humanity” and “[human] civilization” interchangeably and ignore impacts on non-humans.

[2]Geist et al. testify to the long-standing worry of existential catastrophe from nuclear winter, but find that current stockpiles would probably not directly induce one. It remains possible that a nuclear war will induce one by other means (e.g. by greatly increasing these stockpiles).

[3]Less relevantly, a large literature studies the willingness to pay to reduce catastrophic risk where the catastrophe is (or can be modeled as) a negative consumption shock. See especially Barro (2006, 2009), Martin and Pindyck (2015, 2021), Aurland-Bredesen (2019), Weitzman (2009), and Acemoglu and Lensman (2024). Note that following a negative consumption shock, the marginal utility of consumption rises, whereas following an existential catastrophe, it falls to zero. Note also that the latter can happen at most once.

[4]Throughout, we will use the term “growth” as shorthand for “technological development”. We will not discuss other sources of consumption growth.

climate feedback loop, would seem to render the "hazard rate" (probability of existential catastrophe per unit time) positive; and even if it is not, this may change. We present a model in which stagnation is not necessarily safe, and argue from it that the risk-minimizing growth rate is typically positive and often high.

We assume throughout that technology is the only source of x-risk: i.e. that in the absence of an anthropogenic x-risk, we will enjoy a long and flourishing future.[5] Accounting for the possibility of natural x-risks which technology can mitigate would strengthen the headline result.

Two clarifications. First: we make no normative claims about the optimal speed of technological development all things considered, only positive claims about the impact of speeding it on the probability of existential catastrophe. Appendix A offers an argument that those with low discount rates should primarily care about minimizing this probability. But for modeling purposes, the key feature of existential catastrophe is not its normative significance but the fact that it can occur at most once.

Second: we follow Jones (2016, 2024) and many others in modeling technology as "one-dimensional". We do this because we are analyzing the impact of speeding or slowing technological development, not directing it. It may be that some technologies raise x-risk, such as biological weaponry; others lower it, such as vaccination; and the best way to decrease x-risk is to delay the former and speed the latter (Bostrom, 2002). Granting this, we still face the question of whether, *on a given path through the space of technology states*, it is riskier to move more quickly. Existing work assumes it always is; we argue it is often not. This is important because some interventions may primarily change the rate of growth but not its direction, e.g. by affecting R&D subsidies or the rate of population growth. Furthermore, many technologists predict that AI will itself soon accelerate technological development across the board (Grace et al., 2024). If so, efforts to lower x-risk by slowing the development of dangerous AI capabilities[6] may do the opposite on balance unless sufficiently targeted.

Outline. Section 2 considers what follows if the hazard rate is a function only of

[5] In practice, this will presumably entail succumbing to a natural existential catastrophe instead (see Appendix A). From very-long-run historical data on large-scale natural catastrophes, and the typical survival rate of other mammal species, Snyder-Beattie et al. (2019) estimate that the hazard rate from natural x-risk is below one in 870,000 per year.

[6] See e.g. the Future of Life Institute's 2023 call for an AI pause (Future of Life Institute, 2023).

the technology state, or of both the technology state and a stock that accrues over time (e.g. greenhouse gas emissions). Here, stagnation is safe only when the current state is perfectly safe. Otherwise, if future technology states will be safe (perfectly or asymptotically), it is safest to grow as *quickly* as possible; and if not, catastrophe is inevitable whatever the growth rate. Given a positive hazard rate, therefore, faster growth is always weakly safer—regardless of whether technologies on the immediate horizon would raise the hazard rate or lower it.

We then consider two mechanisms through which faster growth can increase risk despite the above.

In Section 3, we suppose that the hazard rate depends not only on the technology state but also directly on the growth rate. That is, the risk of catastrophe in a given year depends not only on the technologies that exist that year—say, the ongoing risk that nuclear weaponry, biotechnology, etc. are used with catastrophic consequences—but also on the number of *experiments* performed that year to develop new technologies. Consider Jones's (2016) analogy between technological development and Russian roulette. We call the first source of risk "state risk" and the second "transition risk".

Accelerating growth has no effect on transition risk if the risk posed by a given experiment is independent of how many experiments happen concurrently, as assumed e.g. by Jones (2016, 2024). Suppose that the future contains a sequence of experiments, each of which will pose some x-risk. Permanent stagnation can lower transition risk by avoiding advanced experiments altogether, but an acceleration that only pulls forward their date leaves cumulative risk unchanged. If the hazard rate is strictly convex in the rate of experimentation, however, then faster growth increases transition risk. The tradeoff between lowering state risk and raising transition risk can render the risk-minimizing growth rate finite, but as long as there is any state risk at all, it remains positive.

In Section 4, we suppose that the hazard rate depends not only on the technology state but also on a policy decision to sacrifice consumption for safety. If policy responds "optimally" to the technology path—in the sense of maximizing expected discounted utility, for an arbitrary discount rate—then the conclusion that faster growth is safer is actually strengthened. This is for two reasons. First, when technology is more advanced, society is richer, so optimal policy is more stringent. The logic is closely analogous to that of Jones (2016, 2024) and to the "environmental Kuznets curve" of Stokey (1998): when consumption is high, the value of life is high and the value of marginal consumption is low.[7] Thus faster growth now may lower x-risk by speeding the arrival not only of safer technology, as in Section 2, but also of safer policy. Second, the value of life is higher, and so optimal policy is more stringent, when *subsequent* growth is expected to be faster, even before consumption has yet risen.

Given policy frictions, the risk-minimizing growth rate may again be finite. If policy cannot mitigate risks as effectively when the technological landscape is changing more rapidly, then speed is risky, as in the case of "pure" transition risk. This effect, if it is strong enough, can on some margins outweigh the contributions of growth to safety outlined above. Still, unless we have reached perfect safety, the risk-minimizing growth rate remains positive.

Section 5 summarizes these results and their limitations.

2 State risk

2.1 State risk only

The hazard rate. The "hazard rate" δ_t is the flow probability at t of (technological) existential catastrophe. In this section we posit that it is an arbitrary non-negative, continuous function of a state variable A_t:

$$\delta_t = \delta(A_t), \quad \delta(\cdot) \geq 0.$$

We will refer to the state variable as "technology", acknowledging the view that technological developments, broadly construed, are the primary drivers of changes in the hazard rate. In this model, therefore, we proceed through a sequence of technology states. A given state may have both risk-inducing features, such as a widespread ability to engineer pathogens, and risk-mitigating features, such as the ability to easily detect novel diseases, develop vaccines, or implement quarantines. If the "technologies" developed over the period after a state A_t on balance raise the hazard rate, $\delta(A_{t+1}) > \delta(A_t)$. If on balance they lower it, $\delta(A_{t+1}) < \delta(A_t)$.

[7] Like these sources, we find that, given a concave enough utility function, enrichment motivates large reallocations from consumption to safety. Our analysis differs in that none of these sources study the conditions under which the probability of a binary event (here, existential catastrophe) is less than 1, nor the risk-minimizing path of a hazard rate over time more generally.

Survival. The probability that we survive to date t is given by

$$S_t \equiv e^{-\int_0^t \delta_\tau d\tau} \quad \Longleftrightarrow \quad \dot{S}_t = -\delta_t S_t, \quad S_0 = 1. \tag{1}$$

The probability that we avoid a catastrophe and enjoy a very long future is

$$S_\infty \equiv \lim_{t\to\infty} S_t = e^{-X}, \tag{2}$$
$$\text{where} \quad X \equiv \int_0^\infty \delta_\tau d\tau.$$

We will refer to $\{\delta_t\}_{t=0}^\infty$ as the *hazard curve*, to the area under the hazard curve X as *cumulative risk*, to $\{S_t\}_{t=0}^\infty$ as the *survival curve*, and to S_∞ as the *probability of survival*.

Note that the probability of survival decreases in cumulative risk, and survival is possible ($S_\infty > 0$) iff cumulative risk is finite. Survival is possible only if the world is on track to eventually be safe, exactly or asymptotically.

2.2 Acceleration

Technology paths and risk density. Let $a \equiv \{a_t\}_{t=0}^\infty$ denote a particular *technology path*, so that on this path, $A_t = a_t$. Unless otherwise stated, we assume that a technology path has a continuous and positive derivative.[8] We denote $a_\infty \equiv \lim_{t\to\infty} a_t$.

On path a, technology crosses every value from a_0 to a_∞ exactly once. So the area under the hazard curve can be found by integrating with respect to technology:

$$X(a) \equiv \int_0^\infty \delta(a_t)dt = \int_{a_0}^{a_\infty} \delta(A)\frac{dt}{dA}dA = \int_{a_0}^{a_\infty} x(A, \dot{a}_A)dA, \tag{3}$$

where

- as shorthand, $\dot{a}_A \equiv \dot{a}_{t^{-1}(A)}$ denotes technology growth per unit time on path a when the technology state is A.
- $x(A, \dot{a}_A)$ denotes the *risk density* at technology state A:
  risk per unit time in A, which in this section equals $\delta(A)$
  $\times$ **"length of time" spent in** A, dt/dA $(\equiv 1/\dot{a}_A)$.

[8] Because technology has not yet been given any substantive interpretation, this just amounts to an indexing of technology states.

Accelerations. Given a technology path a, choose $\underline{A}, \overline{A}$ with

$$a_0 \leq \underline{A} < \overline{A} < a_\infty. \tag{4}$$

Call technology path $\hat{a}$ an *acceleration* to a from $\underline{A}$ to $\overline{A}$ if $\hat{a}_0 = a_0$,

$$\dot{\hat{a}}_A \begin{cases} = \dot{a}_A, & A \in \left(a_0, \underline{A}\right); \\ > \dot{a}_A, & A \in \left(\underline{A}, \overline{A}\right); \\ = \dot{a}_A, & A \in \left(\overline{A}, a_\infty\right), \end{cases} \tag{5}$$

and $\hat{a}$ is continuous, and C^1 except at $\{\underline{A}, \overline{A}\}$. Because risk density at A decreases in $\dot{a}_A$, acceleration weakly lowers the risk endured across the range of technology levels:

$$\begin{aligned} X(\hat{a}) &= X(a) + \Delta X(\hat{a}, a), \\ \text{where} \quad \Delta X(\hat{a}, a) &\equiv \int_{\underline{A}}^{\overline{A}} \left(x(A, \dot{\hat{a}}_A) - x(A, \dot{a}_A)\right) dA \\ &= \int_{\underline{A}}^{\overline{A}} \delta(A)\left(1/\dot{\hat{a}}_A - 1/\dot{a}_A\right) dA \leq 0, \end{aligned} \tag{6}$$

with the inequality strict unless $\delta(A) = 0$ for $A \in [\underline{A}, \overline{A}]$.

This leaves two possibilities. If $X(a)$ is finite, the acceleration decreases cumulative risk by (6) and weakly increases the probability of survival. If $X(a)$ is infinite, the probability of survival is zero with or without the acceleration.[9]

Since an acceleration from $\underline{A}$ temporarily increases the hazard rate if $\delta(\cdot)$ is increasing around $\underline{A}$ (as in Fig. 1a or b), it may appear to contemporaries that the acceleration decreases the probability of survival. Here, however, that is impossible.

In summary:

Proposition 1 (Risk impact of acceleration given state risk only)**.**
Given a technology path "a" and an acceleration "$\hat{a}$", $\Delta X(\hat{a}, a) \leq 0$.
Thus acceleration is always risk-minimizing.

[9] ΔX is finite by the continuity of δ in A, and the continuity of $\dot{a}$ and $\dot{\hat{a}}$ in t and of t in A.

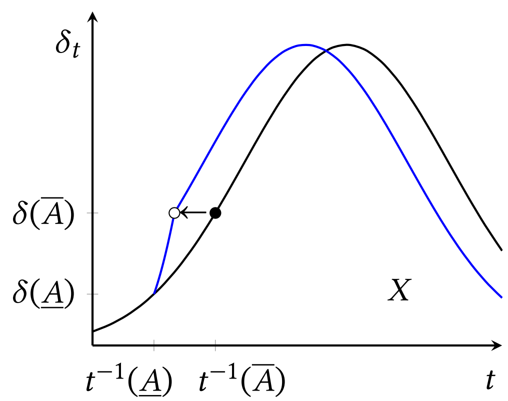

Figure 1a: $X < \infty$;
acceleration lowers X

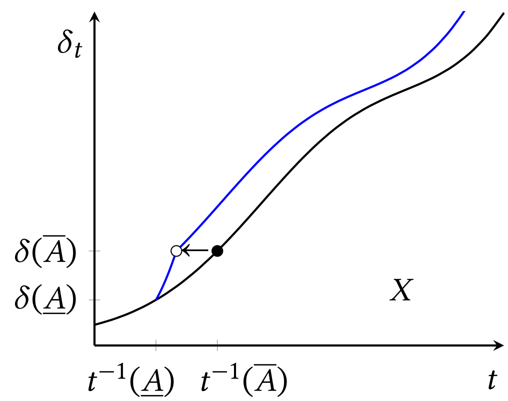

Figure 1b: $X = \infty$;
acceleration has no effect on X

Stagnation. Choosing t^*, and denoting $A^* \equiv a_{t^*}$, call $\hat{a}$ a *stagnation* at A^* if

$$\hat{a}_t = \begin{cases} a_t, & t < t^*; \\ A^*, & t \geq t^*. \end{cases}$$

Stagnations are in some sense extreme decelerations, but their risk impact depends on the value of $\delta(A^*)$.

If $\delta(A^*) > 0$, stagnation at A^* renders cumulative risk infinite. The hazard rate is permanently positive, and survival is impossible. For illustration, consider the implications of a large negative shock today returning the world to the technology state it inhabited in 1925. This reset would doom us to relive the nuclear standoffs, emissions-intensive industrializations, and biotechnological hazards of the past. If any of these pose any existential risk at all, then with enough replays of the past century, catastrophe is inevitable.

Stagnation at A^* is safe only if $\delta(A^*) = 0$, so that from t^* onward, cumulative risk is zero. The key difference between stagnation and mere deceleration is that given deceleration, technology still crosses every state from a_0 to a_∞ once: we simply spend longer in each state and so endure more risk in it. Given stagnation, on the other hand, we avoid states $A > A^*$ altogether.

2.3 Accrued state risk

A generalized model of state-based risk is suggested by the climate modeling literature.

Suppose that, as we spend time in a given technology state, we accrue some stock M on which the hazard rate depends. In state A, the stock grows at rate $m(A)$, so that

$$M_t = M_0 + \int_0^t m(A_\tau)dt,$$

where $m(\cdot) \geq 0$. The hazard rate at t weakly increases in M_t, but also depends on how our technology exacerbates or mitigates the hazard this stock poses:

$$\delta_t = \delta(A_t)\, p(M_t),$$

where $p(\cdot)$ is non-decreasing and continuous and $\delta(\cdot)$, as in the simple state risk model, is non-negative and continuous.[10]

We will illustrate the idea here in the climate change and arms race contexts. Note also that Kasirzadeh (2025) argues that x-risk from AI might take roughly the above form: advanced AI of a given description might not produce a catastrophe immediately, but might introduce sources of social instability which progressively raise the hazard rate.

Climate change. M might denote the quantity of greenhouse gases in the atmosphere, weighted by their contribution to warming. Then $m(A)$ denotes the emissions rate in state A. If the temperature increases logarithmically in atmospheric greenhouse gas concentration,[11] and the probability per unit time of triggering a catastrophic climate feedback loop increases quadratically in the temperature above the preindustrial baseline,[12] we have

$$\delta_t \propto \delta(A_t)\Big(\ln\Big(1 + \int_0^t m(A_\tau)d\tau\Big)\Big)^2,$$

where time 0 denotes the beginning of industrialization.

[10]Note that if $m(\cdot) = 0$ or $p(\cdot)$ is constant, this model reduces to the simple state risk model.

[11]Following Romps et al. (2022).

[12]Roughly following the conventional assumption, from e.g. Nordhaus's DICE model, that damages increase quadratically in temperature above baseline (Nordhaus and Sztorc, 2013).

Technological arms race. Alternatively, M might denote an adversary's military capacity. The rate at which this capacity develops in a given period may depend arbitrarily on the state of one's own technology A (by $m(\cdot)$), e.g. due to the adversary's capacity for espionage or one's own capacity to frustrate it. The probability of an existential catastrophe (or the annihilation of one's own country) increases in M (by $p(\cdot)$) but, importantly, may increase or decrease in A itself (by $\delta(\cdot)$).

For our purposes, the implications of accrued state risk are the same as the implications of simple state risk. Cumulative risk on technology path a equals

$$X(a) = \int_{a_0}^{a_\infty} \delta(A)\, p\Big(M_0 + \int_{a_0}^{A} m(B)\dot{a}_B^{-1}\, dB\Big)\, \dot{a}_A^{-1}\, dA.$$

An acceleration from $\underline{A}$ to $\overline{A}$ weakly lowers cumulative risk both directly, by decreasing the time spent in each technology state $A \in (\underline{A}, \overline{A})$ (raising the $\dot{a}_A$ term in the outer integral), and indirectly, by decreasing the accrual during $(\underline{A}, \overline{A})$ and thus decreasing the hazardous stock in states $A > \underline{A}$ (raising the $\dot{a}_B$ term in the inner integral).[13]

For simplicity, we will work with the simple state risk model going forward.

3 Transition risk

A hazard function of the form $\delta(A)$ captures what we have called "state risk". But risk may also be "transitional": posed by the process of developing and deploying new technologies, rather than by their existence once deployed. This is the intuition captured by Jones's (2016) "Russian roulette" model of technological development and (2024) model of AI risk, and by Bostrom's (2019) analogy to drawing potentially destructive balls from an urn.

We will first consider the case in which all risk is transitional. In this case, stagnation is safe. Nevertheless, even here, we will see that acceleration from a positive-growth baseline may lower or not impact cumulative risk, depending on the elasticity of the hazard rate to the speed of technological development.

[13] If $m(\cdot)$ may be negative, e.g. because in some states carbon is removed from the atmosphere, the risk impact of acceleration is ambiguous in such states. The impact on X via decreasing time spent in each state is negative, but the impact via affecting the stock accrued in later states is positive. In states A with $m(A) \geq 0$, acceleration remains risk-minimizing.

We will then consider the case in which we face both state and transition risk, and characterize risk-minimizing growth when the risks posed by faster growth trade off against the safety that comes from escaping existing risks more quickly.

3.1 Transition risk only

To consider the case in which technological development is the only source of risk, posit that the hazard rate takes the form

$$\delta(A, \dot{A}) = f(A)\dot{A}^{\gamma}, \quad \gamma > 0 \tag{7}$$

where $f(\cdot)$ is positive and continuous.

The $f(A)$ term appears in the hazard function because the safety of the "experiments" needed to develop technologies just beyond the frontier A may depend on what this frontier is. Introducing one new technology in a given period ($\dot{A} = 1$) poses greater risk the further advanced the technology frontier is if $f(\cdot)$ is increasing, and less risk if $f(\cdot)$ is decreasing.[14]

If $\gamma > 1$, a sequence of experiments poses more risk if they are performed in parallel than if they are performed in sequence. This may happen, for example, if society is resilient enough to withstand a sequence of small disasters but not to withstand many simultaneously. If $\gamma < 1$, the experiments pose less risk if performed in parallel.

Consider the case of $f(A) \propto 1/A$ and $\gamma = 1$:

$$\delta_t \propto \dot{A}_t / A_t.$$

Here, each proportional increase to A induces the same hazard, independently of how quickly it occurs. This model is essentially equivalent to the "Russian roulette" model of Jones (2016) and the AI risk model of Jones (2024).

Acceleration and risk. Let a denote a technology path maintaining the conditions listed in Section 2.2.

The impact of acceleration on cumulative risk depends on whether γ is greater or

[14] Alternatively, to interpret one "new technology" as a *proportional* increase to A, rewrite (7) as $\delta = \tilde{f}(A_t)(\dot{A}_t/A_t)^{\gamma}$ where $\tilde{f}(A) \equiv f(A)A^{\gamma}$. On this interpretation, developing more advanced technology poses greater risk iff $\tilde{f}(A)$ increases in A.

less than 1. This can again be seen by integrating the hazard curve with respect to A:

$$X(a) = \int_0^\infty f(a_t)\dot{a}_t^\gamma \, dt = \int_{a_0}^{a_\infty} x(A, \dot{a}_A)\, dA, \quad \text{where} \quad x(A, \dot{a}_A) = f(A)\dot{a}_A^{\gamma-1}.$$

As in (6), given an acceleration $\hat{a}$ from $\underline{A}$ to $\overline{A}$, $\Delta X(\hat{a}, a) = \int_{\underline{A}}^{\overline{A}} \big(x(A, \dot{\hat{a}}_A) - x(A, \dot{a}_A)\big)\, dA$. Observe that the integral is negative if $\gamma < 1$, zero if $\gamma = 1$, and positive if $\gamma > 1$.

Because the Jones models implicitly adopt $\gamma = 1$, they imply that the speed of technological development does not affect cumulative risk, except in that stagnation ($\dot{A} = 0$) eliminates risk entirely.[15]

3.2 State risk and transition risk

Suppose we face both risk types, so that

$$\delta(A, \dot{A}) = h(A) + f(A)\dot{A}^\gamma. \tag{8}$$

Assume that $h(\cdot)$ and $f(\cdot)$ are continuous. Assume also that $h(\cdot)$ and $f(\cdot)$ are strictly positive, to avoid the trivial case in which stagnation is perfectly safe, and that $\gamma > 1$, since we have shown that in the $\gamma \leq 1$ cases acceleration is always risk-minimizing. We now face the tradeoff that acceleration lowers state risk but raises transition risk. We will see that a positive but finite growth rate is risk-minimizing.

The risk-minimizing growth rate. Here, the risk density at state A equals

$$\delta(A, \dot{a}_A)/\dot{a}_A = h(A)\dot{a}_A^{-1} + f(A)\dot{a}_A^{\gamma-1}, \tag{9}$$

which is minimized by

$$\dot{a}_A^* \equiv \left(\frac{1}{\gamma - 1}\frac{h(A)}{f(A)}\right)^{1/\gamma}. \tag{10}$$

Since the indexing of technology states is still arbitrary, we can without loss of generality normalize the risk-minimizing growth rate at a given technology state A, $\dot{a}_A^*$,

[15] This is true even though the models imply a finite technology level A^* at which it is welfare-maximizing to halt technological development: the speed at which we grow to A^* does not affect cumulative risk.

to 1. Under this normalization, transition hazard at the risk-minimizing growth rate equals $f(A)$, and (10) shows that the risk-minimizing ratio of state to transition hazard is $\gamma - 1$. That is, slower growth is safer if state hazard is less than $\gamma - 1$ times as high as transition hazard, and vice-versa. This follows straightforwardly from the fact that the elasticity of transition hazard to the growth rate is $\gamma - 1$ times the negative (unit) elasticity of state hazard to the growth rate. If $\gamma = 2$, for instance, the risk-minimizing growth rate sets the state and transition hazards equal.

Consider a cyclist beside a busy road, with some positive probability per unit time of being hit at any given speed (including zero). Even if the probability of an accident per unit time increases in the cycling speed, halting is not safe: it guarantees that an accident will occur eventually. Indeed, unless the hazard rate more than doubles as speed doubles, at some margin, the safest plan is to bike home as quickly as possible. If the hazard rate does increase superlinearly with speed, the risk-minimizing speed is such that moving 1% more quickly would produce a 1% increase to the hazard rate, just offsetting the 1% decrease in time spent on the road.

The risk-minimizing technology path. Given A_0, let a^* denote the technology path that satisfies (10) at all $A \geq A_0$. As we can see, $\dot{a}^*_A$ rises with $h(A)/f(A)$. That is, risk-minimizing growth accelerates with time if, when the technology state advances, state hazard rises by a greater proportion (or falls by a smaller proportion) than transition hazard does at any given growth rate.

For illustration, suppose

$$h(A) = \bar{h}A^{\alpha}, \quad f(A) = \bar{f}A^{\zeta} \tag{11}$$

for some $\bar{h} > 0$, $\bar{f} > 0$, and assume $\gamma > 1$. Then

$$\dot{a}^*_A \propto A^{\frac{\alpha-\zeta}{\gamma}},$$

so a^*_t grows power-functionally if $\alpha < \zeta + \gamma$, hyperbolically if the inequality is reversed (!), and exponentially if the terms are equal.[16] Positive state risk ensures that stagnation, or even asymptotic stagnation, is never risk-minimizing.

[16]In the last case, substituting (11) into (10) and dividing both sides by A shows that the risk-minimizing growth rate is $\left((\gamma-1)\bar{f}/\bar{h}\right)^{-1/\gamma}$: e.g. if $\bar{f} = \bar{h}$ and $\gamma = 2$, 100% per year.

Substituting (11) into (10), both into (9), and composing the integral, we have

$$X(a^*) = \left[\bar{h}\left(\frac{1}{\gamma-1}\frac{\bar{h}}{\bar{f}}\right)^{-\frac{1}{\gamma}} + \bar{f}\left(\frac{1}{\gamma-1}\frac{\bar{h}}{\bar{f}}\right)^{\frac{\gamma-1}{\gamma}}\right]\int_{A_0}^{\infty} A^{\frac{\zeta-\alpha+\alpha\gamma}{\gamma}}\,dA.$$

It follows that $X(a^*)$ is finite, and survival feasible, iff the exponent on A in the integral is less than -1: that is, iff $\alpha(\gamma-1)+\gamma+\zeta<0$.[17] Because $\gamma>1$, survival is possible only if α or ζ is negative. Intuitively, to survive, either more advanced technology states must (eventually, at least) carry hazard rates that fall toward zero—in this setting, α must be negative—or we must grow ever more quickly, so that the state hazard endured per state, $h(A)/\dot{a}_A$, diminishes. In the latter case, however, a positive value of ζ implies that the transition hazard increases.

In summary:

Proposition 2 (Risk-minimizing growth given state and transition risk)**.**
Given technology path "a", acceleration "â", and hazard function

$$\delta(A,\dot{A}) = h(A) + f(A)\dot{A}^{\gamma}, \quad \gamma>0$$

with continuous $h(\cdot)\geq 0$ and $f(\cdot)>0$:

1. *If $\gamma<1$ $[=1]$, $x(A,\dot{a}_A)$ [weakly] decreases in $\dot{a}_A$, so $\Delta X(\hat{a},a)\leq 0$. Thus acceleration is always risk-minimizing.*
2. *If $\gamma>1$, $x(A,\dot{a}_A)$ is uniquely minimized by $\dot{a}_A^* \equiv \left(\frac{1}{\gamma-1}\frac{h(A)}{f(A)}\right)^{1/\gamma}$. This is positive iff $h(A)>0$.*

Corollary 2.1.
Suppose $h(A)\propto A^{\alpha}$ and $f(A)\propto A^{\zeta}$.

- *Survival is feasible iff $\alpha(\gamma-1)+\gamma+\zeta<0$.*

If this holds:

- *If $\gamma\leq 1$, acceleration always reduces risk.*
- *If $\gamma>1$, the risk-minimizing growth path a^* satisfies $\dot{a}_A^*\propto A^{\frac{\alpha-\zeta}{\gamma}}$.*

[17] This condition is independent of the γ vs. $\alpha-\zeta$ condition for the functional form of $a^*(\cdot)$.

4 Policy

We have assumed so far that cumulative risk depends only on the technology path. More precisely, we have considered the risk impact of speeding or slowing our movement along a sequence of states, where the "state space" has been defined finely enough that our location in it and rate of motion through it captures every feature of the world relevant to the hazard rate.

If we describe the risk-relevant state of the world by a pair of features (A, B), we can say nothing from first principles about the risk impact of accelerating the A-path from a to $\hat{a}$ in isolation. As an important example, suppose the hazard rate depends on the state of technology A and policy B, with

$$\delta(A, B) = A/B,$$

and our policy framework for mitigating risk improves exogenously:

$$B_t = e^{g_B t}, \quad g_B > 0.$$

Then trivially, if A cannot decrease, technological stagnation is safest. Indeed, if $a_t = e^{g_A t}$ for $g_A \in [0, g_B)$, the hazard rate declines exponentially, so X is finite and survival is possible; whereas given a permanent acceleration to $\hat{a}(t) = e^{\hat{g}_A t}$ for $\hat{g}_A \geq g_B$, the hazard rate is constant or rises exponentially, so X is infinite and survival impossible.

In this section we explore the risk-minimizing technology path, a^*, when the hazard rate depends on the state of technology and policy. Instead of an exogenous policy path as above, however, we assume that policy is set by a planner aiming to maximize discounted expected utility. More precisely, we index technology states so that A equals feasible consumption per capita.[18] In each period, the planner decides how much consumption to forego to lower the hazard rate.

As we will see, when policy is set optimally, the conclusion of Section 2—that a faster rate of technological development carries lower cumulative risk—is not only maintained but strengthened. We then consider how policy frictions, making policy less effective (or more costly) when technology changes more quickly, may reintroduce a kind of transition risk.

[18]So the assumption that a_t increases is now substantive: potential consumption grows over time.

4.1 Environment

Preferences. A planner seeks to maximize discounted expected utility,[19] where flow utility is isoelastic in consumption:

$$\int_0^\infty e^{-\rho t} S_t\, u(C_t)\, dt, \quad \rho > 0; \tag{12}$$

$$u(C) = \frac{C^{1-\eta} - 1}{1 - \eta}, \quad \eta > 1. \tag{13}$$

The utility of death is normalized to 0 and the death-equivalent consumption level to 1. We skip the $\eta < 1$ case for two reasons (and the $\eta = 1$ edge case for simplicity).[20]

First, it does not seem to be empirically relevant, either currently (see e.g. Hall (1988) and Lucas (1994)) or, especially, in the very long run (see Appendix A).

Second, if $\eta > 1$, marginal utility in consumption diminishes quickly enough that, given any choice between accelerating consumption growth and increasing the probability of survival, the non-discounted benefits of the latter predominate in the long run (see Appendix A). By contrast, if $\eta < 1$ and consumption grows at rate g, flow utility grows at rate $(1 - \eta)g$ as consumption grows large, and accelerating consumption growth and reducing x-risk both produce streams of increases to expected flow utility that grow indefinitely at rate $(1 - \eta)g$. Thus concern for the long-term future would not generally motivate severely slowing consumption growth for safety in the first place.

Production technology. Index technology states A by potential consumption. Index policy $B_t \in [0, 1]$ by the fraction of potential consumption that is forgone. Consumption at t then equals

$$C_t = A_t(1 - B_t). \tag{14}$$

We will call B the *safety share*, but choices of $B > 0$ may constitute explicit spending on services like pandemic monitoring and/or bans on risky production processes.

[19] We may suppose that the population is fixed and (12) is the expected utility of a representative household, or that population grows exponentially at rate $n < \rho$, the discount rate is $\rho + n$, and the planner adopts the total utilitarian social welfare function.

[20] As is not uncommon in the economics literature on catastrophic risk: e.g. Martin and Pindyck (2015, 2021) impose $\eta > 1$.

We assume that a technology path a satisfies $a_0 \geq 1$.

The hazard rate. The hazard rate is a flexible function of the technology and policy variables. For $A \geq 1$, the hazard function $\delta(A, B)$ is C^2;[21]

D1 decreases and is convex in B, with $\delta(A, 1) = 0$; and

D2 satisfies $\alpha(A, B) < \beta(A, B)$,

where $\alpha(A, B)$ denotes the elasticity of δ to A (which may have any sign) and $\beta(A, B)$ denotes the elasticity of δ to $1 - B$ (which is non-negative).

D1 ensures that the hazard rate is positive if the safety share is not maximal. If a safety share less than 1 can secure perfect safety, the result of Section 4.2—that growth can yield safety by motivating more stringent policy—is strengthened. D1 also ensures that there are weakly diminishing returns to safety spending.

D2 ensures that when technology advances, it is feasible to lower the hazard rate by retaining the former consumption level, allocating all marginal productive capacity to safety measures. That is, if A increases by, say, 1% and $1 - B$ falls by 1%, so that by (14) C stays fixed, the hazard rate falls. If D2 fails (indefinitely), survival is impossible unless consumption is driven to zero: an existential catastrophe by other means.

Note that $\delta(\cdot)$ allows the effectiveness of safety spending to depend arbitrarily on the technology state.

A planner chooses a policy path $b = \{B_t\}_{t=0}^{\infty}$ to maximize discounted expected utility (12)–(13) subject to the budget constraint (14) and a technology path and hazard function.

4.2 The existential risk Kuznets curve

Preliminaries. Given technology path a, let b_a denote the optimal continuous policy path. (Its existence and uniqueness are proved in Prop. 3.)

Let $S(a, b)$ denote the survival curve given technology and policy paths a, b. Define $X(a, b)$ and the hazard curve $\delta(a, b)$ likewise.

Let $v_t(a, b)$ denote the *expected value of the future of civilization*, as of t (assuming

[21] We define $\frac{d\delta}{dy}(A, 0) \equiv \lim_{B \to 0} \frac{d\delta}{dy}(A, B)$ for $y \in \{A, B\}$, and allow these derivatives to be infinite.

survival to t), given survival curve $S(a, b)$ and consumption path $a(1 - b)$:

$$v_t(a, b) \equiv \int_t^{\infty} e^{-\rho(\tau - t)} \frac{S_\tau(a, b)}{S_t(a, b)} u\big(a_\tau(1 - b_\tau)\big) d\tau. \tag{15}$$

Denote $S(a) \equiv S(a, b_a)$, and $v(a)$, $X(a)$, and $\delta_t(a)$ likewise.

Denote the *consumption share* [path] $\tilde{B} \equiv 1 - B$, $\tilde{b} \equiv 1 - b$.

Observation 1. $v_t(a) > 0$.

Proof. It is feasible for the planner to choose $b_t = 0 \ \forall t$. This implements a path of flow utility given survival that begins non-negative (with $C_0 = a_0 \geq 1$) and rises. □

Observation 2. If $\hat{a}_\tau \geq a_\tau$ for all $\tau \geq t$, with strict inequality for some $\tau \geq t$, then $v_t(\hat{a}) > v_t(a)$.

For a proof, see Appendix B.1. For intuition, define policy path b (from t onward) by

$$\hat{a}_t \tilde{b}_t = a_t \tilde{b}_{at}, \tag{16}$$

so the consumption path from t given $\hat{a}, b$ equals that given a, b_a. By D2, we have $\delta_\tau(\hat{a}, b) \leq \delta_\tau(a, b_a)$ for all $\tau > t$, with strict inequality for some $\tau > t$. Thus $\hat{a}$ allows for weakly more consumption and safety than a.

Observation 3. $v_t(a)$ increases to a finite limit.

Proof. Given $\tau > 0$, define $\hat{a}_t \equiv a_{t+\tau}$. Because a increases, $\hat{a}_t > a_t$. By Obs. 2, $v_{t+\tau}(a)[\equiv v_t(\hat{a})] > v_t(a)$. Because $u(C) < \frac{1}{\eta - 1}$, $v_t < \bar{v} \equiv \frac{1}{\rho(\eta - 1)}$. □

Observation 4. Optimal policy at t must satisfy the first-order condition that the loss in flow utility from marginally increasing the safety share weakly exceeds the benefit via reducing the hazard rate and increasing the probability that v_t is realized:

$$\frac{d}{db_{at}} u\big(a_t(1 - b_{at})\big) - \left[\frac{d}{db_{at}} \delta(a_t, b_{at})\right] v_t(a) \leq 0, \tag{17}$$

with inequality only if the marginal value of safety spending is negative even at the $B = 0$ corner. The lower Inada condition on $u(\cdot)$ ensures that $B = 1$ is never optimal. A proof of the necessity of FOC (17) is given in the proof of Prop. 3 (Appendix B.3.)

Example: constant elasticities. Suppose the technology path is $a_t = e^{gt}$ for $g > 0$.

Suppose that the hazard function features constant α and β:

$$\delta(A, B) = \bar{\delta} A^{\alpha}(1 - B)^{\beta}, \quad \bar{\delta} > 0, \quad \beta > \alpha \geq 0, \quad \beta \geq 1, \tag{18}$$

so that the hazard rate falls from $\bar{\delta}A^{\alpha}$ to 0 as policy grows more stringent.

$\beta \geq 1$ maintains D1 and $\beta > \alpha$ maintains D2. We here also impose $\alpha \geq 0$ so that fixing $B < 1$, δ increases in A. This grants that the direct impact of technological development is to weakly increase the hazard rate, against the indirect impact of potentially lowering the hazard rate by motivating more safety spending. If $\alpha < 0$, the hazard rate falls due to technological development alone, as would be necessary for survival in the policy-free model of Section 2.

Substituting (13) and (18) into (17), differentiating, and rearranging:

$$\tilde{b}_{at} = \min\left(\left(\bar{\delta}\beta v_t(a)\right)^{-\frac{1}{\beta+\eta-1}} a_t^{-\frac{\alpha+\eta-1}{\beta+\eta-1}}, 1\right). \tag{19}$$

Recalling that a_t and $v_t(a)$ increase, the optimal policy is to spend nothing on safety until the first term of the maximum is positive; $\tilde{b}_a$ then falls from 1 toward 0.

The reason why has been understood at least since Hall and Jones (2007): when $\eta > 1$, safety is a luxury good. As A rises, if the consumption share $\tilde{B}$ is fixed, the safety benefits of marginally lowering it are valued more highly, because "life" (here, civilization) is more valuable.[22] On the other hand, the utility cost of a proportional consumption cut falls given $\eta > 1$.

Let t_a denote the last period at which zero safety spending is optimal.

Initial risk increases — At $t < t_a$, the hazard rate equals $\delta_t = \bar{\delta}a_t^{\alpha}$ and grows at rate[23]

$$g_{\delta t} = \alpha g \geq 0. \tag{20}$$

[22] Here, also because the hazard rate is higher when A rises and its elasticity in $\tilde{B}$ is constant.

[23] We let g_{xt} denote the exponential growth rate of variable x at t.

Eventual risk declines — After t_a, by (19), the consumption share $\tilde{B}$ grows at rate

$$g_{\tilde{b}_a,t} = -\frac{1}{\beta+\eta-1} g_{v(a),t} - \frac{\alpha+\eta-1}{\beta+\eta-1} g < 0. \tag{21}$$

The hazard rate in turn grows as

$$\begin{aligned} g_{\delta t} &= \alpha g + \beta g_{\tilde{b}_a,t} \\ &= -\frac{(\beta-\alpha)(\eta-1)}{\beta+\eta-1} g - \frac{\beta}{\beta+\eta-1} g_{v(a),t} < 0, \end{aligned} \tag{22}$$

which is negative by $\beta > \alpha$, $\eta > 1$, and Obs. 3. The indirect negative impact of growth on the hazard rate, by increasing the safety share, outweighs any positive impact imposed by α.

The boundedness of v (Obs. 3) gives us the asymptotic negative growth rates of $\tilde{B}$ and δ, as well as that of consumption $C = A\tilde{B}$:[24]

$$\lim_{t\to\infty} g_{Ct} = \left(1 - \frac{\alpha+\eta-1}{\beta+\eta-1}\right) g = \frac{\beta-\alpha}{\beta+\eta-1} g > 0. \tag{23}$$

Survival — Because $\beta > \alpha$, by (22) the hazard rate falls exponentially in the limit. So $X(a) < \infty$ and $S_\infty(a) > 0$.

General result. Increases in productive capacity motivate increases to the “safety share” B under conventional assumptions which imply that safety is a luxury good. Furthermore, our example illustrates that if α and β are fixed and technology grows exponentially, the rise in the safety share renders the probability of survival positive whenever survival is compatible with non-negative consumption growth ($\beta > \alpha$).

However, the “Kuznetsian” dynamic is not strong enough to produce a positive probability of survival in general. We now characterize whether a given hazard function and technology path permit survival, given a planner with preferences (12)–(13), in close to full generality. Though the condition is somewhat complex, it offers a helpful way to evaluate this key property of a hazard function. It also illustrates why, given slow growth or low risk aversion (η), the planner may sometimes choose a policy that

[24] The fact that v rises to an upper bound does not strictly imply $g_v \to 0$. A proof that $g_v \to 0$ in this example is available upon request.

precludes survival despite its feasibility.

Proposition 3 (The existential risk Kuznets curve and survival)**.**
Given a hazard function $\delta(\cdot)$, a technology path "a" that is either C^1 with a positive derivative or an acceleration to one that is, and preferences (12)–(13):

1. *An optimal continuous policy path b_a exists and is unique.*

Define $\bar{a}(p) \equiv \lim_{t\to\infty} a_t t^{-\frac{p}{\eta-1}}$ and

$$D(k) \equiv \begin{cases} \lim_{t\to\infty} \left[-\frac{d\delta}{dB}\left(a_t, 1 - t^{\frac{k}{\eta-1}} / a_t\right)\right] t^{\frac{k\eta}{\eta-1}} / a_t, & \lim_{p\to 1^+} \bar{a}(p) > 0; \\ \lim_{t\to\infty} \left[-\frac{d\delta}{dB}\left(a_t, 0\right)\right] t^k, & \bar{a}(1) = 0. \end{cases}$$

2a. *If $\lim_{k\to 1^+} D(k) = 0$, then $S_\infty(a) > 0$.*

2b. *If $D(1) > 0$ and $\beta(\cdot)$ is upper-bounded, then $S_\infty(a) = 0$.*

Proof. See Appendix B.3. Note that the result does not cover cases in which the $D(k)$ limit is undefined at or near 1. □

To interpret the survival condition, recall that if the optimal consumption share $(1 - b_{at})$ is interior in the limit, the flow utility gained by marginally raising it $a_t u'(C_t)$ must equal the cost, via increased risk, of marginally raising it. Rearranging (17):

$$\left(a_t(1 - b_{at})\right)^{1-\eta} \tag{24}$$
$$= \frac{d\delta(a_t, b_{at})}{d(1 - b_{at})}(1 - b_{at})v_t(a). \tag{25}$$

Suppose that $v_t = \bar{v}$ is constant (c.f. Obs. 3). Suppose also that $\beta(\cdot)$ is constant:

$$(1 - b_t)\frac{d\delta(a_t, b_t)}{d(1 - b_t)} = \beta\, \delta_t(a, b). \tag{26}$$

Substituting (26) into (25), we see that if the optimal safety share is eventually interior, the integral of the hazard curve $X(a)$ converges and $S_\infty > 0$ if (24) is bounded above in the limit by t^{-k} for $k > 1$. If (24) is bounded below in the limit by a curve proportional to t^{-1}, $X(a) = \infty$ and $S_\infty = 0$.

Let $b_{kt} \equiv 1 - t^{\frac{k}{\eta-1}} / a_t$ denote the policy path maintaining $(a_t(1 - b_{kt}))^{1-\eta} = t^{-k}$. If a_t grows faster than $t^{\frac{p}{\eta-1}}$ for some $p > 1$, $D(k)$ is proportional to the limit of (25)/(24)

with b_k in place of b_a. If there is a $k \in (1, p)$ with $D(k) = 0$, then even policy path b_k—which is feasible and permits survival—lowers the hazard rate too slowly. If $D(1) > 0$, even b_1—which does not—lowers the hazard rate too fast.

If $\bar{a}(1) = 0$, a grows so slowly that b_k is infeasible (i.e. eventually exceeds 1) for any $k > 1$. We are thus functionally in the state-risk-only case of Section 2. Unless $\beta(\cdot)$ can grow arbitrarily high, so that small safety expenditures grow arbitrarily effective, survival requires technology growth eventually to lower the hazard rate faster than $1/t$ on its own.

4.3 Illustration

Example 1: constant elasticities. The optimal policy path and the corresponding hazard curve are simulated below, for hazard function (18), technology path

$$a_t = 2e^{gt}, \tag{27}$$

and the parameter values in Table 1.

ρ	0.02	$\bar{\delta}$	0.00012
η	1.5	α	1
g	0.02	β	2

Table 1: Simulation parameters for Figure 2

The values of ρ, η, and g have been chosen as central estimates from the macroeconomics literature. $a_0 = 2$ is chosen so that the value of a statistical life-year (VSLY) at $t = 75$ is 4x consumption per capita, roughly matching Klenow et al. (2025).[25] That is, the simulation might be taken to begin in 1950, with year 75 being the present. $\bar{\delta}$, α, and β are chosen so that the present hazard rate is ~0.1%, matching Stern's (2007) oft-cited figure; the hazard rate begins to fall at $t \approx 100$; and the growth and decay of the hazard rate are non-negligible.

Example 2: a lower Inada condition on safety spending. As (20) and (22) show,

[25]They estimate that this ratio was approximately 5 in the United States in 2019. The figure must be adjusted upward in light of economic growth since 2019, but downward insofar as the model is intended to describe the path of optimal policy across all countries advanced enough to be deploying existentially hazardous technology.

an arguably unrealistic feature of hazard function (18) is that as soon as it is worth spending on safety at all, optimal spending rises rapidly enough that the hazard rate falls. This can be modified without altering the intuition of the example by using a hazard function with a lower Inada condition on safety spending, such as

$$\delta(A, B) = \bar{\delta} A^{\alpha}(1-B)^{\beta-1}(1-B^{\epsilon}), \quad \epsilon \in \left(\frac{1}{2}, 1\right). \tag{28}$$

We choose $\epsilon = 0.6$ and the parameter values of Table 1. The lower Inada condition ensures that the optimal safety share b_a is always positive, and as it rises smoothly, the hazard rate rises and falls smoothly.[26]

On these parameters, the probability of survival S_{∞} from $t = 75$ onward is ~69%.

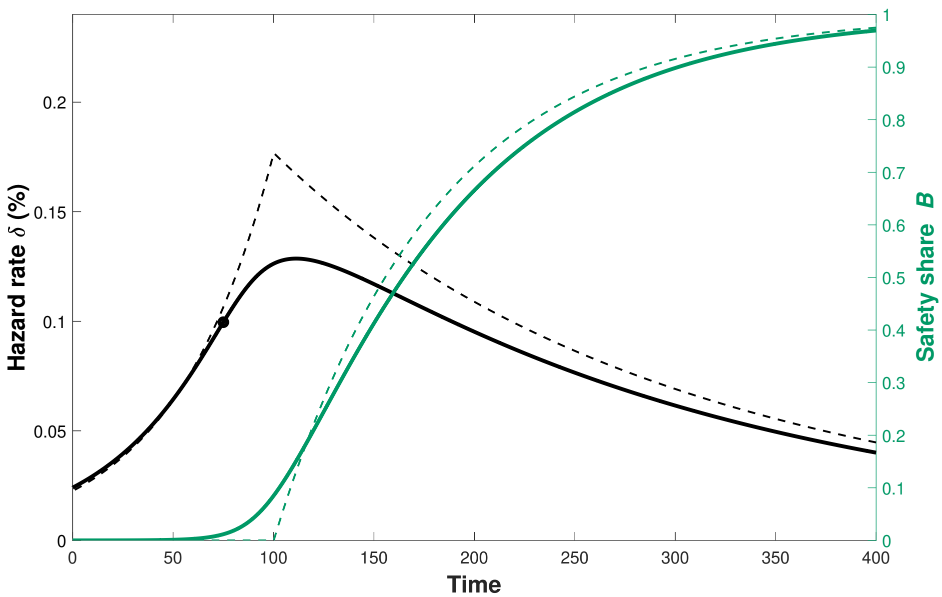

Figure 2: Evolution of policy and risk given hazard functions (18) (dashed) and (28) (solid)

Derivations and code for replicating the simulations may be found in Appendix C.

4.4 Acceleration

If the policy choice at t depended only on the technology state at t—if we had $b_{at} = b(a_t)$—then, given optimal policy, the hazard function could be expressed as a function

[26]Note that (18) is the special case of (28) with $\epsilon = 1$. Iff $\epsilon < 1/2$, the hazard rate *always* falls instead of exhibiting a Kuznets curve. When $b_t \approx 0$, from (61) we have $\dot{b}_t \approx K b_t$ (for some $K > 0$); and so from (62), $\dot{\delta}_t$ has the same sign as $\alpha g - \epsilon K b_t^{2\epsilon-1}$.

of A. The impact of acceleration on cumulative risk would thus be precisely as in Section 2, stated in Prop. 1.

Here, with $b_{aA} \equiv b_{a,t^{-1}(A)}$ denoting the optimal safety share on technology path a when the technology state equals A, cumulative risk equals

$$X(a) \equiv \int_{a_0}^{a_\infty} x(A, b_{aA}, \dot{a}_A)dA, \quad \text{where} \quad x(A, b_{aA}, \dot{a}_A) = \delta(A, b_{aA})/\dot{a}_A.$$

Define $v_A(a)$ analogously to b_{aA}. By (17), b_{aA} depends not only on A but also on $v_A(a)$. In particular, because $\delta(\cdot)$ in convex in B (D1) and is C^2,[27]

$$b_{aA} \text{ is continuous and weakly increasing in } v_A(a). \tag{29}$$

When the future is more valuable, at a given technology state, it is worth spending more to save it, except perhaps in the $b_{aA} = 0$ corner solution.

Let $\hat{a}$ be an acceleration to a from $\underline{A}$ to $\overline{A}$.[28]

Observation 5. $v_A(\hat{a}) > v_A(a)\ \forall A < \overline{A}$.

For a proof, see Appendix B.2. The intuition is straightforward: when one is in a given state, a faster-growing future is more valuable, because (by Obs. 2) it allows for more consumption and more safety.

Acceleration thus lowers cumulative risk not only by shrinking the time spent at each technology state, as in Prop. 1, but also potentially by motivating more stringent policy at each state before the acceleration ends. Assuming that the acceleration is not anticipated before $\underline{A}$, we have:

$$\begin{aligned}
\Delta X(\hat{a}, a) &\equiv \int_{\underline{A}}^{\overline{A}} \left(x(A, b_{\hat{a}A}, \dot{\hat{a}}_A) - x(A, b_{aA}, \dot{a}_A)\right) dA \\
&= \underbrace{\int_{\underline{A}}^{\overline{A}} \left(\delta(A, b_{aA})/\dot{\hat{a}}_A - \delta(A, b_{aA})/\dot{a}_A\right) dA}_{<0 \text{ as in Prop. 1: } \delta>0,\ \dot{\hat{a}}_A > \dot{a}_A.} + \underbrace{\int_{\underline{A}}^{\overline{A}} \left(\delta(A, b_{\hat{a}A})/\dot{\hat{a}}_A - \delta(A, b_{aA})/\dot{\hat{a}}_A\right) dA.}_{\leq 0:\ v_A(\hat{a})>v_A(a) \text{ (Obs. 5)} \implies b_{\hat{a}A} \geq b_{aA} \text{ (29)}.}
\end{aligned}$$

(An anticipated acceleration would motivate more stringent policy even before $\underline{A}$.)
In summary:

[27] I.e. its derivative is C^1, so the continuity of b_{aA} in $v_A(a)$ follows from the implicit function theorem.
[28] Note that if a grows exponentially, $\hat{a}$ is a level effect.

Proposition 4 (Risk impact of acceleration given optimal policy)**.**
Given a technology path "a" and an acceleration "â", $\Delta X(\hat{a}, a) < 0$.
Thus acceleration is always risk-minimizing.

Furthermore, acceleration lowers cumulative risk by weakly more than in the absence of a policy response.

Comparing this to Prop. 1, $\Delta X(\hat{a}, a)$ is here strictly negative only because we have assumed $\delta(A, B) > 0$ (unless $B = 1$, which never obtains). More substantively, optimal policy strengthens the tendency for acceleration to lower state risk for two reasons.

1. Whereas a state-risk-only model is agnostic about whether later states *will* be safer, policy introduces a tendency in this direction: when consumption grows, the utility cost of marginally sacrificing consumption falls and the value of life rises, often quickly enough to permit survival (Prop. 3).

2. The prospect of *future* increases to consumption growth lowers the *present* hazard rate, because when the value of the future is greater, it is worth sacrificing more today to prevent its ruin (Prop. 4).

With reference to Fig. 1, the first implication of optimal policy is that X is more likely finite, and the second is that the hazard rate decreases in anticipated future growth.

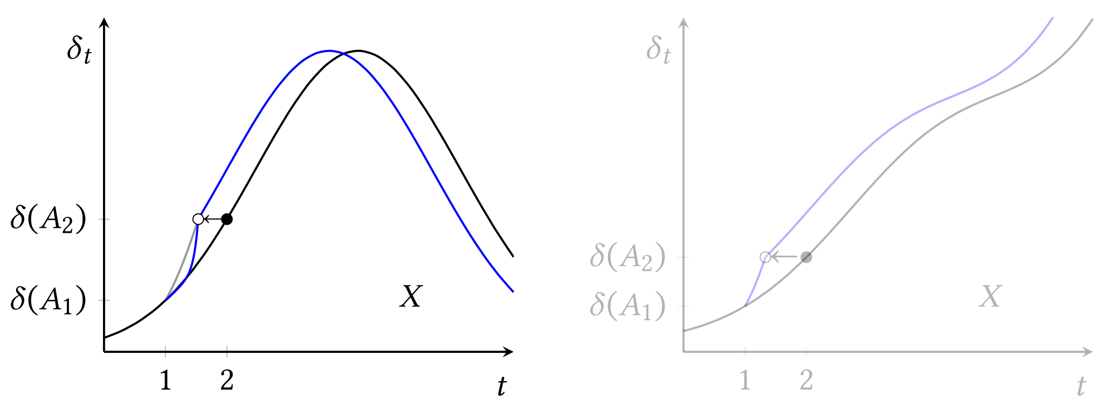

Figure 3: Optimal policy (i) facilitates finite X (the left graph rather than the right) and (ii) lowers the hazard rate associated with each technology level during an acceleration (the gap between the blue and gray lines from $\delta(A_1)$ to $\delta(A_2)$)

4.5 Policy frictions

We have assumed so far that as the technology state changes, policy can frictionlessly reallocate resources in the way that best balances consumption and safety. This has shown that lack of concern for the future is not enough to overturn the positive relationship between growth and safety: for any value of ρ, if the planner can always maintain her ideal resource allocation, giving her more resources by accelerating growth lowers cumulative risk in the long run.

However, this frictionlessness is unrealistic. When technology changes more quickly, safety regulations and expenditures may not be appropriate to the threats of the day.[29] Indeed, this is a primary motivation for positing that the hazard rate increases in the speed of technological change, as explored in Section 3. Suppose therefore that the hazard rate is a function of technology A and *effective* safety spending B_{eff}, where B_{eff} increases in B but decreases in $\dot{A}$.

Analogy to transition risk — We will model the risks of faster growth, via less effective safety spending, in terms cleanly comparable to the reduced-form analysis of transition risk in Section 3. Recalling that $B \in [0, 1]$, consider the possibilities

$$B_{\text{eff}} = B^{1+m(A)\dot{A}^{\gamma}}, \tag{30}$$

$$B_{\text{eff}} = Be^{-m(A)\dot{A}^{\gamma}}, \tag{31}$$

where arbitrary $m(A) > 0$ allows the effect of rapidly introducing some technology on the contemporaneous effectiveness of safety spending to depend on the technology in question. In (30), the effective safety share ranges from 0 to 1 as B does, such that in principle allocating all economic activity to safety efforts would eliminate risk. In (31), a positive growth rate upper-bounds B_{eff} below 1, and so introduces some risk that safety spending cannot eliminate.

Then consider the hazard function

$$\delta(A, B_{\text{eff}}) = h(A) \ln(1/B_{\text{eff}}), \tag{32}$$

where the elasticity of $h(\cdot)$ is bounded below 1 to satisfy D2. Substituting (30) and (31)

[29] See Shulman and Thornley (2024) and Jones (2025), who argue that the current policy response to existential risk is far from optimal even under a relatively high discount rate.

into (32), we have, letting $f(A) \equiv h(A)m(A)$,

$$\delta = \big(h(A) + f(A)\dot{A}^{\gamma}\big)\ln(1/B), \tag{33}$$

$$\delta = h(A)\ln(1/B) + f(A)\dot{A}^{\gamma}. \tag{34}$$

Fixing $B \in (0, 1)$, each case reduces to the hazard function of Section 3.2.

If we drop the "1" in the exponent of (30), we drop the $h(A)$ term from (33) and reproduce the transition-risk-only hazard function of Section 3.1.

Acceleration — Let a be a technology path and $\hat{a}$ be an acceleration to it. If $b_{\hat{a}} = b_a$, it follows from the above that $\Delta X(\hat{a}, a) < 0$ or > 0 under the same conditions as in Prop. 2, and the risk-minimizing growth path a^* is as characterized there.[30] Here, however, the policy path depends on the technology path. We will see that faster growth *tends* to increase safety spending, as in Section 4.4, making the risk-minimizing growth path faster than characterized in Prop. 2; but that here in general the effect is ambiguous.

By the first-order condition (17), noting that (32) exhibits a lower Inada condition on safety spending and thus that safety spending is always interior, we have

$$\begin{aligned}
\frac{du}{db_{at}} &= \frac{d\delta}{db_{at}} v_t(a), \\
\implies b_{at}\big(1 - b_{at}\big)^{-\eta} &= a_t^{\eta-1}\big(h(a_t) + f(a_t)\dot{a}_t^{\gamma}\big)v_t(a), \\
b_{aA}\big(1 - b_{aA}\big)^{-\eta} &= A^{\eta-1}\big(h(A) + f(A)\dot{a}_A^{\gamma}\big)v_A(a) \quad \text{in case (30)}, && (35) \\
\implies b_{aA}(1 - b_{aA})^{-\eta} &= A^{\eta-1}h(A)v_A(a) \quad \text{in case (31)}. && (36)
\end{aligned}$$

Thus $b_{\hat{a}A}$ may differ from b_{aA} for two reasons: because $\dot{\hat{a}}_A^{\gamma} > \dot{a}_A^{\gamma}$ (in case (30) only) or because $v_A(\hat{a}) \neq v_A(a)$.

Fixing v_A, faster growth motivates more safety spending in case (30): in riskier situations, safety efforts tend to be prioritized more highly because they accomplish more in absolute terms. In case (31), however, speeding growth has no direct effect on policy: the fact that mitigating risk is more difficult fully offsets the fact that there is more risk to mitigate.

Fixing $\dot{a}_A$, here, unlike in Section 4.2 (Obs. 2), faster growth after A has an ambiguous effect on v_A. This is because here, when growth is faster, it is more costly to

[30] As detailed throughout the rest of this section, the existence of a policy response will tend to make the $X(a) < \infty$ case more likely, so that inducing $\Delta X < 0$ raises S_∞ instead of having no effect.

achieve a given degree of safety. Indeed, if $\gamma > 1$, a catastrophe at A is guaranteed as $\dot{a}_A \to \infty$ for any $B < 1$ (even if $B = 1$, in case (31)). For moderate accelerations, however, the $v_A(\hat{a}) > v_A(a)$ case, in which the planner prefers (marginally) faster growth, is presumably more empirically relevant. States generally subsidize R&D, not tax it.

5 Conclusion

Technologies can pose or mitigate existential risks. Stagnation is safe, as assumed in existing literature, only if the current technology state poses no such risks. Otherwise, for any fixed direction of technological development, safety requires growth, and perhaps rapid growth. The conventional wisdom that slower is safer holds only if policy frictions, or risks posed directly by the process of technological development, are sufficiently severe.

	State risk			**Transition risk**		
	Alone	Frictionless policy		Alone	With state risk	With state risk; due to frictions
$\uparrow A \implies \downarrow\downarrow \delta$	∞	∞	$\gamma < 1$	∞	∞	∞
$\uparrow A \not\Rightarrow \downarrow\downarrow \delta$; strict policy effective + desirable	any	∞	$\gamma = 1$	any	∞	∞
$\uparrow A \not\Rightarrow \downarrow\downarrow \delta$; strict policy ineffective or undesirable	any	any	$\gamma > 1$	0	finite	typically faster than $\leftarrow$

Table 2: Summary of risk-minimizing growth rates

We omit the "accrued state risk" case of Section 2.3, as it behaves largely like simple state risk, and the "Transition risk" table omits cases in which survival is impossible. Recall that given both risk types and $\gamma > 1$, the finite risk-minimizing growth rate may rise to ∞ or fall to 0, depending on how the relative contributions of state and transition risk evolve.

Even if technological development to date has raised the hazard rate on balance, and will do so in the immediate future, the tendency for safety to be a luxury good suggests that x-risk is likely to exhibit a Kuznets curve. That is, we may indeed be in Sagan's (1997) "time of perils" (see Appendix A). If so, securing safety today comes with a massive long-term benefit. Even putting aside the consumption benefits of faster growth, however, the safety benefit of slower (and thus perhaps less disruptive or better regulated) technological development trades off against that of escaping the perils more quickly.

This is not an argument against ever regulating risky technologies. Indeed, an important way technological development can lower cumulative risk is by hastening the day when regulation is strict. Some reactions to calls for heavy AI regulation, e.g. by Andreessen (2023), might be read as arguing that our "safety share" should never be high. If that is so, it is not for reasons presented in this paper.

Our framework highlights that for those interested in reducing cumulative existential risk, quantifying the relative contributions of state and transition risk, and forecasting how these will evolve, would be valuable. A more precise understanding of the policy distortions around the regulation of risky technologies would be particularly valuable, both for determining whether they are severe enough to contribute significantly to transition risk and for determining how responsive policy is likely to be if the hazard rate sharply rises. Slower growth may well be safer. For now, however, our results suggest that even those exclusively concerned with long-term survival should often encourage technological advances despite their short-term hazards, and advocate risk-reduction measures today only when they are sufficiently targeted and the costs to broad-based technological progress are sufficiently small.

References

Acemoglu, Daron and Todd Lensman, "Regulating Transformative Technologies," *American Economic Review: Insights*, 2024, *6* (3), 359–76.

Andreessen, Marc, "The Techno-Optimist Manifesto," 2023. Open letter.

Aurland-Bredesen, Kine Josefine, "The Optimal Economic Management of Catastrophic Risk." PhD dissertation, Norwegian University of Life Sciences School of Economics and Business 2019.

Baranzini, Andrea and François Bourguinion, "Is Sustainable Growth Optimal?," *International Tax and Public Finance*, 1995, *2*, 341–356.

Barro, Robert J., "Rare Disasters and Asset Markets in the Twentieth Century," *The Quarterly Journal of Economics*, 2006, *121* (3), 823–66.

_ , "Rare Disasters, Asset Prices, and Welfare Costs," *American Economic Review*, 2009, *99* (1), 243–64.

Bostrom, Nick, "Existential Risks: Analyzing Human Extinction Scenarios," *Journal of Evolution and Technology*, March 2002, *9* (1), 1–35.

_ , "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas*, November 2003, *15* (3), 1–35.

_ , "The Vulnerable World Hypothesis," *Global Policy*, 2019, *10* (4), 455–476.

Caputo, Michael R., *Foundations of Dynamic Economic Analysis: Optimal Control Theory and Applications*, Cambridge University Press, 2005.

Center for AI Safety, "Statement on AI Risk," 2023.

Christiano, Paul, "What Failure Looks Like," 2019. Blog post, LessWrong.

Farquhar, Sebastian, John Halstead, and Owen Cotton-Barratt, "Existential Risk: Diplomacy and Governance," Technical Report 2017.

Future of Life Institute, "Pause Giant AI Experiments: An Open Letter," March 2023.

Geist, Edward, Alvin Moon, Henry H. Willis, and Anu Narayanan, "Chapter 8. Nuclear War: Summary of Risk," in "Global Catastrophic Risk Assessment," RAND Corporation, 2024.

Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner, "Thousands of Authors on the Future of AI," 2024. Preprint.

Hall, Robert, "Intertemporal Substitution in Consumption," *Journal of Political Economy*, 1988, *96* (2), 339–357.

Hall, Robert E. and Charles I. Jones, "The Value of Life and the Rise in Health Spending," *Quarterly Journal of Economics*, 2007, *122* (1), 39–72.

Hanson, Robin, "Limits To Growth," 2009. Blog post, Overcoming Bias.

Harrod, Roy F., *Towards a Dynamic Economics*, London: Macmillian, 1948.

Jones, Charles I., "Life and Growth," *Journal of Political Economy*, 2016, *124* (2), 539–578.

— , "The A.I. Dilemma: Growth Versus Existential Risk," *American Economic Review: Insights*, 2024, *6* (4), 575–590.

— , "How Much Should We Spend to Reduce A.I.'s Existential Risk?," 2025. Working paper.

Karnofsky, Holden, "This Can't Go On," 2021. Blog post, Cold Takes.

Kasirzadeh, Atoosa, "Two Types of AI Existential Risk: Decisive and Accumulative," *Philosophical Studies*, 2025, *1* (7), 1–29.

Klenow, Peter J., Charles I. Jones, Mark Bils, and Mohamad Adhami, "Population and Welfare: Measuring Growth when Life is Worth Living," 2025. Working paper.

Koopmans, Tjalling C., "On the Concept of Optimal Economic Growth," 1963. Cowles Foundation Discussion Paper 392.

Kulveit, Jan, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud, "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development," 2025. Working paper.

Lucas, Deborah, "Asset Pricing with Undiversifiable Risk and Short Sales Constraints: Deepening the Equity Premium Puzzle," *Journal of Monetary Economics*, 1994, *34* (3), 325–342.

Martin, Ian W. R. and Robert S. Pindyck, "Averting Catastrophes: The Strange Economics of Scylla and Charybdis," *American Economic Review*, 2015, *105* (10), 2947–2985.

— **and** — , "Welfare Costs of Catastrophes: Lost Consumption and Lost Lives," *The Economic Journal*, 2021, *131* (634), 946–969.

Millett, Piers and Andrew Snyder-Beattie, "Existential Risk and Cost-Effective Biosecurity," *Health Security*, 2017, *15*, 373–383.

Méjean, Aurélie, Antonin Pottier, Marc Fleurbaey, and Stéphane Zuber, "Catastrophic Climate Change, Population Ethics and Intergenerational Equity," *Climatic Change*, 2020, *163*, 873–890.

Nordhaus, William, "A Review of the *Stern Review on the Economics of Climate Change*," *Journal of Economic Literature*, 2007, *45* (3), 686–702.

__ , "The Economics of Tail Events with an Application to Climate Change," *Review of Environmental Economics and Policy*, 2011, *5* (2), 240–57.

__ **and Paul Sztorc**, "DICE 2013R: Introduction and User's Manual," April 2013.

Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity*, New York: Bloomsbury, 2020.

__ , "Shaping humanity's longterm trajectory," 2024.

Parfit, Derek, *Reasons and Persons*, Oxford: Oxford University Press, 1984.

Posner, Richard A., *Catastrophe: Risk and Response*, New York: Oxford University Press, 2004.

Romps, David M., Jacob T. Seeley, and Jacob P. Edman, "Why the Forcing from Carbon Dioxide Scales as the Logarithm of Its Concentration," *Journal of Climate*, 2022, *35* (13), 4027–47.

Sagan, Carl, *Pale Blue Dot: A Vision of the Human Future in Space*, Ballantine Books, 1997.

Shulman, Carl and Elliott Thornley, "How Much Should Governments Pay to Prevent Catastrophes? Longtermism's Limited Role," in Jacob Barrett, Hilary Greaves, and David Thorstad, eds., *Essays on Longtermism*, Oxford: Oxford University Press, 2024.

Snyder-Beattie, Andrew E., Toby Ord, and Michael B. Bonsall, "An Upper Bound for the Background Rate of Human Extinction," *Scientific Reports*, December 2019, *9* (1), 11054.

Solow, Robert M., "The Economics of Resources or the Resources of Economics," *American Economic Review*, 1974, *64* (2), 1–14.

Steffen, Will, Johan Rockström, Katherine Richardson, Timothy M. Lenton, Carl Folke, Diana Liverman, Colin P. Summerhayes, Anthony D. Barnosky, Sarah E. Cornell, Michel Crucifix, Jonathan F. Donges, Ingo Fetzer, Steven J. Lade, Marten Scheffer, Ricarda Winkelmann, and Hans Joachim Schellnhuber, "Trajectories of the Earth System in the Anthropocene," *The Proceedings of the National Academy of Sciences of the United States of America*, 2018, *115* (33), 8252–9.

Stern, Nicholas, *The Economics of Climate Change: The Stern Review*, Cambridge and New York: Cambridge University Press, 2007.

Stokey, Nancy, "Are There Limits to Growth?," *International Economic Review*, 1998, *39* (1), 1–31.

Thorstad, David, "Existential Risk Pessimism and the Time of Perils," 2022. GPI Working Paper Series No. 1-2022.

Weitzman, Martin L., "On Modeling and Interpreting the Economics of Catastrophic Climate Change," *Review of Economics and Statistics*, 2009, *91*, 1–19.

A Why focus on survival?

When making tradeoffs over time, it is uncontroversial to discount later periods for reasons of uncertainty. Whether to include a rate of pure time preference in the social welfare function as well—even across long time horizons involving multiple generations—has been a matter of disagreement at least since the objections of Harrod (1948, p. 40), Koopmans (1963), and Solow (1974). This question is especially central to the debate over optimal climate policy: Nordhaus (2007) prominently argues that pure time preference should be included, Stern (2007) that it should not.

Bostrom (2003) argues that with no pure time preference, welfare-maximizing policy is, to a close approximation, whatever minimizes existential risk. We here formalize his argument by providing simple conditions under which the approximation holds.

Notation. We build on the notation of Section 2. Let

- $\mathcal{A}$ denote the *space of technology states* and A denote a generic technology state;
- a, with $a_t \in \mathcal{A}$ for $t \in [0, \infty)$, denote a *technology path*; and
- $S_t(a)$ for $t \in [0, \infty]$ denote the probability that no anthropogenic existential catastrophe has occurred by t given technology path a.

A *technology state* is a description of the state of human civilization fine-grained enough that (i) the survival curve $\{S_t\}$ depends only on the technology path and (ii) flow utility at time t depends only on the technology state, i.e. $u_t = u(A_t)$. Call a technology path a continuous if $u(a_t)$ is continuous in t.

Suppose that at some (known or unknown) time T, an exogenous natural event will occur which will unavoidably end human civilization if it has not ended already, such as the death of the sun or the heat death of the universe. Discounting only for uncertainty, the expected utility of the future given continuous technology path a and exogenous end-time T equals

$$U(a,T) \equiv \int_0^T S_t(a)u(a_t)dt. \tag{37}$$

Result. A pair of continuous technology paths $a, \hat{a}$ is *asymptotically utility-equivalent* if

$$\lim_{t\to\infty} \frac{u(\hat{a}_t)}{u(a_t)} = 1 \tag{38}$$

and, for some $\underline{t}$, $u(a_t)$ is bounded above 0 across $t > \underline{t}$.[31]

For example, suppose $\lim_{t\to\infty} u(a_t) = \lim_{t\to\infty} u(\hat{a}_t) = \bar{u} > 0$. Perhaps on both paths, the population is constant, consumption per person grows without bound, and flow utility in consumption is bounded above by $\bar{u}$. Then

$$\lim_{t\to\infty} \frac{u(\hat{a}_t)}{u(a_t)} = \frac{\bar{u}}{\bar{u}} = 1.$$

Throughout Section 4—the only section in which utility appears at all—accelerations from a to $\hat{a}$ are always asymptotically utility-equivalent for this reason.[32]

Alternatively, suppose that on either path, individual flow utility approaches the same limit $\bar{u}$, and population eventually grows cubically as we expand into space at some maximum feasible speed; but expansion begins one period earlier on $\hat{a}$ than on a. Then

$$\lim_{t\to\infty} \frac{u(\hat{a}_t)}{u(a_t)} = \lim_{t\to\infty} \frac{(t+1)^3\bar{u}}{t^3\bar{u}} = 1.$$

Proposition 5 (Only survival matters)**.**

[31]Prop. 5 also holds if $u(a_t)$ is asymptotically bounded below zero. In this case the implication is that all that matters in the long run is to *increase* x-risk.

[32]This is because when $\eta > 1$, $\bar{u} = \frac{1}{\eta-1}$. Note that in the $\eta = 1$ (logarithmic) case, flow utility grows linearly like gt if consumption grows exponentially at rate g, so an acceleration from a to $\hat{a}$ is still asymptotically utility-equivalent, by $\lim_{t\to\infty} \frac{g(t+k)}{gt} = 1 \ \forall k$.

If continuous technology paths a, $\hat{a}$ are asymptotically utility-equivalent and $S_\infty(a) > 0$,

$$\lim_{T\to\infty} \frac{U(\hat{a},T)}{U(a,T)} = \frac{S_\infty(\hat{a})}{S_\infty(a)}.$$

Proof. Since $u(a_t)$ is continuous and asymptotically bounded above zero, $\lim_{T\to\infty} U(a,T) = \infty$. By (38), if $\lim_{T\to\infty} U(\hat{a},T) < \infty$, we must have $S_\infty(\hat{a}) = 0$, so the proposition follows immediately. If $\lim_{T\to\infty} U(\hat{a},T) = \infty$, by L'Hôpital's Rule and the fundamental theorem of calculus the limit equals $\lim_{T\to\infty}[S_T(\hat{a})u(\hat{a}_T)]/[S_T(a)u(a_T)]$. $S_\infty(\hat{a})$ is defined by the monotone convergence theorem, and the proposition follows by (38). □

The time of perils. This result is only relevant if T is high enough that, for any pair of paths that might reasonably be under consideration, $U(\hat{a},T)/U(a,T)$ is near its limit. This seems likely for the following reasons.

From very-long-run historical data on large-scale natural catastrophes, and the typical survival rate of other mammal species, Snyder-Beattie et al. (2019) estimate that the hazard rate from natural x-risk is below one in 870,000 per year. Insofar as we, unlike other species, will develop technological solutions to some natural x-risks, we should expect T to be even greater than 870,000.

Karnofsky (2021), building on Hanson (2009), offers an intuitive case that technological development in a welfare-relevant sense cannot continue at anything close to its current pace for over 10,000 more years. This suggests that a very long-term failure to achieve our potential, flow utility must stagnate (or at best grow cubically; see above) well before a natural catastrophe's expected arrival date.

Finally, to maintain that on a path a we can roughly apply the discount factor $S_\infty(a)$ to the entire interval $[0,T]$, we must argue that (i) $S_T(a)$ is non-negligible and (ii) $S_t(a)$ approaches its limit well before T. That is, we must argue that on the technology paths under consideration, humanity may not destroy itself, but if it does, it will probably do so within the next, say, few thousand years. Parfit (1984) called this the view that we live at the "hinge of history", and Sagan (1997) the "time of perils". As both recognized, and as Thorstad (2022) emphasizes, this hypothesis underlies the case for taking survival to be approximately all that matters in our present circumstances. The "existential risk Kuznets curve" we find in Section 4 supports the hypothesis.

B Proofs

B.1 Proof of Observation 2

Suppose $\hat{a}_t \geq a_t$ for all $t \geq 0$, with strict inequality for some t. Define b as in (16), observing that

$$\begin{aligned} b_t &= 1 - \frac{a_t}{\hat{a}_t}(1 - b_{at}) \in [b_{at}, 1], \\ u\big(\hat{a}_t(1 - b_t)\big) &= u\big(a_t(1 - b_{at})\big) \equiv u_t. \end{aligned}$$

Then

$$\begin{aligned} v_0(\hat{a}, b) - v_0(a, b_a) &= \int_0^\infty e^{-\rho t}\big(S_t(\hat{a}, b) - S_t(a, b_a)\big) u_t dt \\ &= \int_0^\infty h(t) f(t) dt; \end{aligned} \tag{39}$$

$$h(t) \equiv \left(\frac{S_t(\hat{a}, b)}{S_t(a, b_a)} - 1\right), \quad f(t) \equiv e^{-\rho t} S_t(a, b_a) u_t.$$

Because by D2 and (16)

$$\delta_t(\hat{a}, b) \leq \delta_t(a, b_a), \tag{40}$$

by (1) we have

$$h(0) = 0, \quad h'(t) \geq 0. \tag{41}$$

By Obs. 1,

$$F(t) \equiv \int_t^\infty f(\tau) d\tau > 0. \tag{42}$$

Integrating (39) by parts, and observing that $f(t) = -F'(t)$, we have

$$v_0(\hat{a}, b) - v_0(a, b_a) = \big[- F(t) h(t) \big]_0^\infty + \int_0^\infty F(t) h'(t) dt. \tag{43}$$

By (41) and (42), the last term is non-negative. By (41), $-F(0)h(0) = 0$. Finally

$$\begin{aligned}&\lim_{t\to\infty} -F(t)h(t)\\ =\ &\lim_{t\to\infty}\left(S_t(a,b_a) - S_t(\hat{a},b)\right)\int_t^{\infty} e^{-\rho\tau}\frac{S_\tau(a,b_a)}{S_t(a,b_a)}u_\tau d\tau.\end{aligned}$$

Since (i) the term outside the integral lies between 0 and 1 in absolute value, (ii) $S_\tau(a,b_a)/S_t(a,b_a) \leq 1$, and (iii) $u_\tau < \frac{1}{\eta-1}$, the limit is zero.

Because $\hat{a}_t > a_t$ for some t, the continuity of technology paths implies that the inequalities of (40) and thus (41) are strict for a positive measure of times. It follows that the last term of (43) is positive.

The proofs for an initial period greater than 0 are precisely analogous.

B.2 Proof of Observation 5

Let $t^{-1}(A)$, $\hat{t}^{-1}(A)$ denote when technology state A is reached on paths a, $\hat{a}$ respectively. Choose $A' < \overline{A}$, let $\Delta t \equiv t^{-1}(A') - \hat{t}^{-1}(A') \geq 0$, and define $\tilde{a}$ by

$$\tilde{a}_t = \hat{a}_{t-\Delta t},$$

so "$\tilde{t}^{-1}(A')$"$=t^{-1}(A')$. Observe that $v_A(\hat{a}) = v_A(\tilde{a})\ \forall A$, so $v_{A'}(\hat{a}) = v_{t^{-1}(A')}(\tilde{a})$. Since $\tilde{a}_\tau \geq a_\tau\ \forall \tau > t^{-1}(A')$, with strict equality for some such τ, we have $v_{t^{-1}(A')}(\tilde{a}) > v_{t^{-1}(A')}(a) \equiv v_{A'}(a)$ by Obs. 2.

B.3 Proof of Proposition 3

Proof of part 1. We will prove that a unique continuous optimal policy path b_a exists for any technology path a that either (i) has a continuous, positive derivative or (ii) is an acceleration to a path that does.

Necessary and sufficient conditions — The planner's optimization problem features one choice variable $\tilde{b}$ and one state S. Expected flow utility at t is $S_t u(a_t\tilde{b}_t)$ for a C^2 function $u(\cdot)$ that is strictly concave and obeys the lower Inada condition. The law of motion for S is $-S_t\delta(a_t, 1-\tilde{b}_t)$ for a C^2 function $\delta(\cdot)$. Because a is independent of $\tilde{b}$, we may treat it as a function of time.

Letting v denote the costate variable on S, the current value Hamiltonian corresponding to the problem is

$$\mathcal{H}(S_t, \tilde{b}_t, v_t, \mu_t, t) = S_t u(a_t \tilde{b}_t) - v_t S_t \delta(a_t, 1 - \tilde{b}_t) + \mu_t(1 - \tilde{b}_t), \tag{44}$$

where μ_t is the the Lagrange multiplier on $\tilde{b}_t$. We impose $\tilde{b}_t \leq 1$ but not $\tilde{b}_t \geq 0$ because the latter can never bind, by the lower Inada condition on $u(\cdot)$.

Equation (44) satisfies the Mangasarian concavity condition that $\mathcal{H}_t$ is weakly concave in S_t and $\tilde{b}_t$. So applying Caputo (2005), Theorems 14.3-4 and Lemma 14.1,[33] given continuous paths of $\tilde{b} \in [0, 1]$ and $S \in [0, 1]$ with $S_0 = 1$ and $\dot{S}_t = -S_t \delta(a_t, \tilde{b}_t)$, we have that the $\tilde{b}, S$ path is optimal if—and, among continuous paths $\tilde{b}$ and S, only if—for some semi-differentiable path of v and some semi-continuous path of $\mu \geq 0$, at all t the first-order and transversality conditions are satisfied:

$$\frac{d\mathcal{H}}{d\tilde{b}_t}(S_t, \tilde{b}_t, v_t, \mu_t, t) = \mu_t \frac{d\mathcal{H}}{d\mu_t}(S_t, \tilde{b}_t, v_t, \mu_t, t) = 0, \quad \frac{d\mathcal{H}}{d\mu_t}(S_t, \tilde{b}_t, v_t, \mu_t, t) \geq 0, \tag{45}$$

$$\lim_{t\to\infty} e^{-\rho t} v_t = \lim_{t\to\infty} e^{-\rho t} v_t S_t = 0. \tag{46}$$

Given paths b, S satisfying the above and corresponding paths v and μ, v_t is continuous and satisfies

$$\dot{v}_t = \rho v_t - \frac{d\mathcal{H}}{dS_t} = \rho v_t - u(a_t \tilde{b}_t) - v_t \dot{S}_t = \big(\rho + \delta(a_t, 1 - \tilde{b}_t)\big) v_t - u(a_t \tilde{b}_t) \tag{47}$$

except at discontinuity points of $\tilde{b}$, where v's right and left derivatives may differ.

The first-order condition — Given a continuous path v, only

$$\tilde{b}_t(v) = \begin{cases} 1, & a_t u'(a_t) - \frac{d\delta}{d\tilde{b}_t}(a_t, 0) v_t \geq 0, \\ \tilde{b}_t \;:\; a_t u'\big(a_t \tilde{b}_t\big) - \frac{d\delta}{d\tilde{b}_t}(a_t, 1 - \tilde{b}_t) v_t = 0, & \text{otherwise}; \end{cases} \tag{48}$$

$$\mu_t(v) = a_t u'(a_t) - \frac{d\delta}{d\tilde{b}_t}(a_t, 1 - \tilde{b}_t) v_t \tag{49}$$

satisfy (45) for all t. The path $\tilde{b}(v)$ is well-defined by the continuous differentiability of

[33] Caputo (2005) uses the more general present value notation. Because the control problem at hand is exponentially discounted, we here use the simpler current value notation.

$\delta(\cdot)$ in $\tilde{b}_t$ and the fact that $u(\cdot)$ and $\delta(\cdot)$ strictly increase in $\tilde{b}_t$, with the former strictly concave and the latter convex. Also, $\tilde{b}(v)$ is right-continuous by the twice continuous differentiability of $u(\cdot)$ and $\delta(\cdot)$, the right-continuity of the right derivative of a, and the implicit function theorem. The path $\mu(v)$ is then also right-continuous by the composition of continuous functions.

To show that there exists an optimal path, and that only one such path is semi-continuous, it will now suffice to show that there is a unique path v for which (46)–(47) are satisfied given $\tilde{b}(v)$ and its implied S path, and that $\tilde{b}(v)$ is continuous.

The transversality condition — The solution to differential equation (47) is

$$v_t = e^{\int_0^t (\rho+\delta_\tau)d\tau}\left(v_0 - \int_0^t e^{-\int_0^\tau (\rho+\delta_q)dq} u(a_\tau \tilde{b}_\tau)d\tau\right) \tag{50}$$

$$\implies v_0 = \int_0^t e^{-\rho\tau} S_\tau u(a_\tau \tilde{b}_\tau)d\tau + e^{-\rho t} S_t v_t. \tag{51}$$

Since (51) is continuous in t (by the boundedness of $u(\cdot)$ and the continuous evolution of S) and holds for all t, v satisfies (46)–(47) iff

$$v_0 = \int_0^\infty e^{-\rho t} S_t u(a_t \tilde{b}_t)dt. \tag{52}$$

Given (48), v_t determines $\tilde{b}_t(v)$ for all t. Given (47), v_t and $\tilde{b}_t$ determine the right derivative of v for all t. Given v_0, therefore, there is a unique path v—and thus $\tilde{b}$, and thus S—compatible with (47)–(48). We will now show that there is at least one value of v_0 for which (52) is satisfied, given the corresponding $\tilde{b}$ and S paths. For such a v_0, the corresponding variable paths by construction satisfy (45)–(46).

Existence — Let $v(v_0)$ and $\tilde{b}(v_0)$ denote the unique paths of v and $\tilde{b}$ compatible with (47)–(48) for which $v_0(v_0) = v_0$. By (50), $\lim_{v_0 \to -\infty} v_t(v_0) = -\infty \; \forall t > 0$. By (48), therefore, for each $t > 0$, there is a $\tilde{v}_0$ such that $\tilde{b}_t(v_0) = 1 \; \forall v_0 < \tilde{v}_0$. Choose $\tau > 0$ and $\tilde{v}_0$ low enough that $v_\tau(\tilde{v}_0) < 0$ and thus $\tilde{b}_\tau(\tilde{v}_0) = 1$. By (47), because $u(a_t \tilde{b}_t) \geq 0$, $\dot{v}_t < 0$. We thus have $v_t(\tilde{v}_0) < 0$, and thus $\tilde{b}_t = 1$, for all $t \geq \tau$.

Observe that if $v_0 < \tilde{v}_0$, $v_t(v_0) < v_t(\tilde{v}_0)$ for all t. Otherwise, by the continuity of v, there would be a t with $v_t(v_0) = v_t(\tilde{v}_0)$, and integrating (47), with (48) substituted for $\tilde{b}_t$, would allow us to identify $v_0 = \tilde{v}_0$. Thus, if $v_0 < \tilde{v}_0$, $\tilde{b}_t(v_0) \geq \tilde{b}_t(\tilde{v}_0) \; \forall t$. It follows that some $\underline{v}_0$ is less than (52) at $\tilde{b} = \tilde{b}(\underline{v}_0)$.

Because (52) is upper-bounded (Obs. 3), some $\bar{v}_0$ exceeds (52) at $\tilde{b} = \tilde{b}(\bar{v}_0)$.

By (48) and the implicit function theorem, $\tilde{b}_t$ is continuous in v_t for all t. (47) then implies that $\dot{v}_t$ is defined and continuous in v_t for all t, and thus that $v_t(v_0)$, then $x_t(v_0)$, and then the right-hand side of (52) are continuous in v_0 for all t. It follows from the intermediate value theorem that (52) holds for some $v_0 \in (\underline{v}_0, \bar{v}_0)$.

Uniqueness — The uniqueness condition of Caputo (2005), Thm. 14.4 does not directly apply because the Hamiltonian is linear, not strictly concave, in S. This can be remedied by defining the state variable as e.g. S^2 without affecting any other results.

Uniqueness (among continuous $\tilde{b}$ paths) also follows from the facts that a path is optimal iff v_0 attains its maximum feasible value and that, given (45)–(46), v_0 determines a unique path for every variable.

Proof of part 2. By first-order condition (17), on $b = b_a$ we have

$$\begin{aligned} &\big(a_t(1-b_t)\big)^{1-\eta} && (53)\\ &\geq \Big[-\frac{d}{db_t}\delta(a_t,b_t)\Big](1-b_t)v_t. && (54)\end{aligned}$$

Fast "a" case — If $\lim_{p\to 1^+}\bar{a}(p) > 0$, there is a $p > 1$ with $\lim_{t\to\infty} a_t t^{-\frac{k}{\eta-1}} > 0$ for $k < p$. For such k, define b_k and the corresponding consumption path $a(1-b_k)$ by

$$\begin{aligned} b_{kt} &= 1 - t^{\frac{k}{\eta-1}}\big/a_t,\\ a_t(1-b_{kt}) &= t^{\frac{k}{\eta-1}}. \end{aligned} \tag{55}$$

a. If $\lim_{k\to 1^+} D(k) = 0$, then for some $k \in (1,p)$, for any $\kappa > 0$, $b_{at} < \bar{v} b_{kt}$ for large t. Choose $\kappa = \bar{v}$ (Obs. 3). Given that $a_t(1-b_{at})$ is lower-bounded in the limit by $\bar{v}\cdot$(55), (53) is upper-bounded in the limit by $\bar{v}^{1-\eta}t^{-k}$ on $b = b_a$. By D1, δ is concave in B and thus in $\tilde{B}$, and $\delta(\cdot,1) = 0$. So for large t we have

$$\bar{v}^{-\eta}t^{-k} > \Big[-\frac{d}{db_{at}}\delta(a_t,b_{at})\Big](1-b_{at}) \equiv \Big[\frac{d}{d\tilde{b}_{at}}\delta(a_t,1-\tilde{b}_{at})\Big]\tilde{b}_{at} \geq \delta_t(a). \tag{56}$$

Thus the integral $X(a)$ is finite and $S_\infty(a) > 0$.

b. If $D(1) > 0$, $a_t(1 - b_{at})$ is upper-bounded in the limit by $\kappa t^{\frac{1}{\eta-1}}$ for some $\kappa > 0$. Thus b_a is interior, and (53–54) holds with equality on $b = b_a$, with both sides lower-bounded in the limit by κt^{-1}. Because for any b

$$\beta(a_t, b_t)\delta(a_t, b_t) = \left[-\frac{d}{db_t}\delta(a_t, b_t) \right](1 - b_t), \tag{57}$$

upper-boundedness of $\beta(\cdot)$ implies that κt^{-1} lower-bounds $\delta_t(a)$ as well.

Slow "a" case — Suppose $\bar{a}(1) = 0$.

a. If $\lim_{k \to 1^+} D(k) = 0$, then for some $k > 1$, $t^{-k}/\bar{v}$ upper-bounds $\delta_t(a, 0)$ as in (56) (with 0 in place of b_a). Because $\delta(\cdot)$ decreases in B, the bound also applies to $\delta_t(a)$.

b. If $D(1) > 0$ and $\beta(\cdot) \leq \bar{\beta}$, then there is a T_0 and $\kappa_0 > 0$ such that, for all $t > T_0$ with $b_{at} = 0$,

$$-\frac{d\delta}{db_{at}}(a_t, 0) > \kappa_0/t.$$

Because (57) holds for all b and all t, we have $\delta_t(a) \geq (\kappa_0/\bar{\beta})/t$ for all $t > T_0$ with $b_{at} = 0$. For t with $b_{at} > 0$, optimality requires (53)=(54). So, in conjunction with (57),

$$\delta_t(a) > a_t^{1-\eta}/\bar{\beta}.$$

By $\bar{a}(1) = 0$, there is a T_1 and $\kappa_1 > 0$ such that for all $t > T_1$ with $b_{at} > 0$, $\delta_t(a) > (\kappa_1/\bar{\beta})/t$.

C Transition dynamics for simulations

For simulating the transition dynamics, it is helpful to find $\dot{\tilde{b}}_{at}$ and $\dot{\delta}_t(a)$ as functions of t and $\tilde{b}_{at}$ in the regime where $\tilde{b}_{at}$ is interior. Since hazard function (18) is the special case of (28) with $\epsilon = 1$, the calculations below apply to both simulations. For simplicity we will drop the "a" arguments and subscripts.

FOC:

$$
\begin{aligned}
\frac{d}{db_t}u\big(a_t(1-b_t)\big) &= \frac{d}{db_t}\delta(a_t, b_t)v_t \\
\implies a_t^{1-\eta}\tilde{b}_t^{-\eta} &= \bar{\delta}a_t^{\alpha}\tilde{b}_t^{\beta-2}\Big((\beta-1)\big(1-(1-\tilde{b}_t)^{\epsilon}\big)+\epsilon\tilde{b}_t(1-\tilde{b}_t)^{\epsilon-1}\Big)v_t \\
\implies v_t &= \frac{1}{\bar{\delta}}\frac{a_t^{1-\eta-\alpha}\tilde{b}_t^{2-\eta-\beta}}{(\beta-1)\big(1-(1-\tilde{b}_t)^{\epsilon}\big)+\epsilon\tilde{b}_t(1-\tilde{b}_t)^{\epsilon-1}} \qquad (58) \\
\implies \dot{v}_t &= v_t\Big((1-\eta-\alpha)g+(2-\eta-\beta)\dot{\tilde{b}}_t/\tilde{b}_t \qquad (59) \\
&\qquad -\epsilon\frac{\beta+(1-\beta-\epsilon)\tilde{b}_t}{(\beta-1)\big((1-\tilde{b}_t)^{1-\epsilon}-1\big)+(\epsilon+\beta)\tilde{b}_t}\frac{\dot{\tilde{b}}_t}{1-\tilde{b}_t}\Big).
\end{aligned}
$$

From the first-order condition with respect to the state variable S_t,

$$
\dot{v}_t = v_t\big(\rho+\delta_t\big)-u\big(a_t\tilde{b}_t\big) = v_t\Big(\rho+\bar{\delta}a_t^{\alpha}\tilde{b}_t^{\beta-1}\big(1-(1-\tilde{b}_t)^{\epsilon}\big)\Big)-\frac{(a_t\tilde{b}_t)^{1-\eta}-1}{1-\eta}. \qquad (60)
$$

Substituting (58) into (59) and (60), setting the results equal, and solving for $\dot{\tilde{b}}_t$ yields

$$
\begin{aligned}
\dot{\tilde{b}}_t = {} & \tilde{b}_t\big((\beta-1)(1-\tilde{b}_t)^{1-\epsilon}-\beta+(\epsilon+\beta)\tilde{b}_t\big)(1-\tilde{b}_t) \\
& \Big((2-\eta-\beta)\big((\beta-1)((1-\tilde{b}_t)^{1-\epsilon}-1)+(\epsilon+\beta)\tilde{b}_t\big)(1-\tilde{b}_t)-\epsilon\big(\beta-(\epsilon+\beta)\tilde{b}_t\big)\tilde{b}_t\Big)^{-1} \\
& \Big(\rho+\bar{\delta}a_t^{\alpha}\tilde{b}_t^{\beta-1}\big(1-(1-\tilde{b}_t)^{\epsilon}\big)-g(1-\alpha-\eta)- \qquad (61) \\
& \frac{(a_t\tilde{b}_t)^{1-\eta}-1}{1-\eta}\bar{\delta}a_t^{\alpha+\eta-1}\tilde{b}_t^{\beta+\eta-2}\big((\beta-1)\big(1-(1-\tilde{b}_t)^{\epsilon}\big)+\epsilon\tilde{b}_t(1-\tilde{b}_t)^{\epsilon-1}\big)\Big).
\end{aligned}
$$

Differentiating the hazard function (28) with respect to t yields

$$
\dot{\delta}_t = \bar{\delta}a_t^{\alpha}\tilde{b}_t^{\beta-1}\big(1-(1-\tilde{b}_t)^{\epsilon}\big)\Big(\alpha g+(\beta-1)\frac{\dot{\tilde{b}}_t}{\tilde{b}_t}+\epsilon\frac{(1-\tilde{b}_t)^{\epsilon-1}}{1-(1-\tilde{b}_t)^{\epsilon}}\dot{\tilde{b}}_t\Big). \qquad (62)
$$

Scripts for replicating Figure 2 using (61) and (62), and the estimate of S_∞ in Example 2, are provided here: https://philiptrammell.com/static/ERAG_code.zip.