Existential Risk and Growth

Leopold Aschenbrenner* and Philip Trammell†

February 11, 2024

Abstract

Technology increases consumption but can create or mitigate "existential risk" to human civilization. In a model of endogenous technology regulation, the willingness to sacrifice consumption for safety grows as the value of life rises and the marginal utility of consumption falls. As a result, in a broad class of cases, when technology is optimally regulated, existential risk follows a Kuznets-style inverted U-shape. This suggests an economic foundation for the prominent view that we are living through a once-in-history "time of perils". Though accelerating technological development during such a period may initially increase risk, it typically decreases cumulative risk in the long run. When technology is regulated optimally, therefore, there is typically no tradeoff between technological progress and the probability of existential catastrophe.

*OpenAI. I am grateful to the Centre for Effective Altruism, the University of Oxford's Future of Humanity Institute, and Columbia ISERP for their support.

†Global Priorities Institute and Department of Economics, University of Oxford. Contact: philip.trammell@economics.ox.ac.uk. Thanks to Danny Bressler, Lennart Stern, and Michael Wiebe for suggesting the idea; to many people at GPI, Ben Snodin, Luis Mota, Tyler Cowen, Pete Klenow, and especially Chad Jones for helpful comments; and to Alex Holness-Tofts and Arvo Muñoz for research assistance on an earlier draft.

"If you are going through hell, keep going."
–Winston Churchill

1 Introduction

Technological progress can bring immense prosperity. Its impact on *existential risk*—the risk of human extinction, or, equivalently for decision purposes, of an equally complete and permanent destruction of human welfare—strikes many as more ambiguous (Bostrom (2002), Posner (2004), Farquhar et al. (2017), Ord (2020), Jones (2023)). Advances in vaccine technology render us less vulnerable to humanity-destroying plagues, for instance; advances in gain-of-function virology arguably make them more likely (Millett and Snyder-Beattie, 2017).

If some existential risks are a permanent byproduct of technologically advanced civilization, an eventual existential catastrophe is inevitable. Civilization will avoid destroying itself technologically only if it pursues a policy of "degrowth": self-destruction by another name.

On the other hand, if, absent a near-term existential catastrophe, we can eventually achieve both advancement and stability, then a near-term existential catastrophe destroys a future that might otherwise have been very valuable and very long. The intuition that we are living through a "time of perils"—a temporary period of high existential risk—was perhaps most famously expressed by Sagan (1997), who coined the phrase, and the prodigious implications for those especially concerned about the long-term future were emphasized early by Parfit (1984) and more recently by Ord (2020). If we are living through a time of perils, risky growth today might buy a short-run increase to human welfare at a severe long-run cost.

This raises the possibility of a tradeoff: concern for the long-run survival of human civilization may motivate slowing or abandoning development, at least outside of sustainability-focused domains such as green energy technology. Sentiments along these lines from an environmentalist perspective go back most notably to the Club of Rome's 1972 report calling for a recognition of the "Limits to Growth", and have recently reemerged with prominent calls to pause AI development (Future of Life Institute, 2023). Jones (2023) explores how to make the tradeoff between AI development and AI risk, under the assumption that such a tradeoff exists.

Is this assumption justified? Would slowing technological development lower existential risk?

We shed light on this question by developing a dynamic model of the tradeoff between consumption and existential safety.

We begin in Section 2 by introducing an economic environment in which the technological frontier grows exogenously and the *hazard rate*—the flow probability of an existential catastrophe—is an especially simple function of the technology level and policy choices. As new potentially dangerous technologies are introduced, a planner, discounting the future at a positive rate, decides how much potential consumption to sacrifice for the sake of lowering the hazard rate.

In Section 3 we find that, in the specified environment, the chosen policy path generates an "existential risk Kuznets curve". That is, the hazard rate rises and then falls with time. Early in time, when the expected discounted value of the future of civilization is relatively low and the marginal utility of consumption is high, it is worthwhile to adopt risky technologies as they arrive, tolerating increases to the hazard rate for the sake of growing consumption rapidly. Later, when the expected value of the future is higher and the marginal utility of consumption has fallen, substantial risk mitigation becomes worthwhile.

This insight mirrors the logic of Stokey (1998) and Brock and Taylor (2005), on which environmental damages rise and then fall with economic development, and of Jones (2016), on which growth yields increases in the value of life relative to marginal consumption. It appears briefly in Cotton-Barratt (2015), who notes that efforts to reduce existential risk today may be more valuable than efforts to do so in the future, because safety efforts will be better funded "in a wealthier world". This dynamic provides a natural economic foundation for the view that we may indeed be living through a once-in-history time of perils. Under the simple functional forms of Sections 2–3 for growth, preferences, and risk, the hazard rate ultimately falls toward zero quickly enough that the probability of escaping the time of perils is positive.

Our model of catastrophic risk differs importantly from those of Martin and Pindyck (2015, 2021) and Aurland-Bredesen (2019). That literature studies a society's willingness to pay to reduce the risk of catastrophes that are, or are essentially equivalent to, proportional consumption cuts. In such a context there are no wealth effects: the fraction of consumption one is willing to sacrifice to avoid a proportional consumption cut is, by definition, independent of one's baseline level of consumption. As emphasized by Jones (2016), and as noted above, reductions in risks to life are luxury goods, given standard preferences: as consumption rises, the marginal utility of risk-reduction rises, whereas the marginal utility offered by other consumption goods falls. Wealth effects therefore play a central role in the existential-risk-focused model studied here.

Section 4 details the implications of the model for the impact of a shock to technology growth on existential risk. The impact of a temporary increase to the technology level is intuitive. Early in time, when technological progress is associated with increases to the hazard rate, temporarily advancing the technological frontier raises risk; late in time, doing so lowers risk. Importantly, however, the effect of a permanent level or growth effect is always to speed civilization through the time of perils. Though such effects may temporarily raise the hazard rate, therefore, they ultimately lower it enough to lower the cumulative probability of a catastrophe.

In short, while the intuition of a time of perils may have a compelling economic foundation, it is a foundation that largely undermines the case for slowing technological development out of concern for long-run safety. For accelerating technological development to increase cumulative risk there must be a sufficiently severe and lasting regulatory inefficiency, not merely insufficient concern by regulators for the long-run future.

This analysis might be compared with that of Baranzini and Bourguinion (1995). In a model in which growth can pose existential risk, Baranzini and Bourguinion define a growth path to be "sustainable" if it (a) minimizes the probability that an anthropogenic existential catastrophe ever occurs and (b) features non-decreasing consumption given survival. They then find conditions under which the optimal growth path, in the conventional sense of maximizing expected discounted utility, is sustainable. We do something like the reverse: we find conditions under which technological advances, when regulated with a view to maximizing expected discounted utility, lower the probability of an anthropogenic existential catastrophe.

The two central ingredients of the model of Sections 2–4 are the technology growth path and the model of the hazard rate. In Sections 5 and 6, we test the robustness of the conclusions above by generalizing both, finding simple conditions that are sufficient for accelerations to technology growth to increase safety in the long run. We find that sustained growth is compatible with long-term survival under some arguably plausible assumptions and incompatible under others. However, we also find that when an efficiently regulated path does allow for survival, the central lessons of Section 4 about the negative relationship between growth and existential risk are typically maintained.

Section 7 concludes by discussing the limitations of this analysis and the value of further research on the relationship between existential risk and growth.

2 The economic environment

2.1 Technology

The maximum feasible level of consumption at t equals the technology level A_t. Actual consumption is A_t multiplied by a policy choice $x_t \in [0, 1]$:

$$C_t = A_t x_t. \tag{1}$$

The tradeoff at the heart of this paper is that a technologically advanced civilization can risk self-destruction, and that this risk can be lowered at some cost to consumption, as represented here by a choice of x below 1. (We denote the choice variable x to remind the reader that higher choices of x come with higher existential risk.) Choices of x below 1 may constitute bans on the adoption of consumption-increasing but risky production processes, and/or allocations of resources to the production of safety goods and services, like pandemic monitoring. The relationship we assume between A, x, and the degree of existential risk, in the baseline model, is given below.

The technology frontier A grows at a positive, exogenous, constant exponential rate g:

$$\dot{A}_t = A_t g, \quad g > 0, A_0 > 0.$$

Alternative growth paths are explored in Section 5.

2.2 Hazard rate

A time-varying hazard rate δ_t represents the flow probability of anthropogenic existential catastrophe. δ_t is a function of the technology level A_t and the policy choice x_t, and is increasing in x_t. For now, we will assume that the elasticities of the hazard rate in A and in x are constant, so that the hazard function equals

$$\delta(A_t, x_t) = \bar{\delta} A_t^{\alpha} x_t^{\beta}, \quad \bar{\delta} > 0,\ \beta > \alpha > 0,\ \beta > 1. \tag{2}$$

We impose $\beta > \alpha > 0$ and $\beta > 1$ to satisfy three desiderata.[1]

The first is that, fixing $x_t > 0$, δ_t increase in A_t. In the context of hazard function (2), this of course requires that $\alpha > 0$. The assumption that δ_t increases in A_t is necessary if we are to concede the assessment that the development of hazardous technologies has rendered an anthropogenic existential catastrophe more likely now than it was centuries ago. The proportion $1 - x$ of potential consumption sacrificed for the sake of existential safety has only increased alongside technological development: having once been zero, it is a small but positive share today.[2] If it had remained fixed, the hazard rate would presumably have followed a weakly higher path.

Second, the elasticity of δ_t with respect to x_t is assumed to exceed the elasticity of δ_t with respect to A_t; i.e., $\beta > \alpha$. This is equivalent to the condition that, when technology advances, it is always feasible to lower the risk level by retaining the former consumption level, allocating all marginal productive capacity to existential safety measures. This may be seen by substituting $x_t = C_t/A_t$ (from (1)) into the hazard function (2), yielding

$$\delta_t = \bar{\delta} A_t^{\alpha-\beta} C_t^{\beta}.$$

Fixing C, the hazard rate falls over time iff $\beta > \alpha$. If it is (indefinitely) infeasible to lower the hazard rate while fixing consumption, as it is in this

[1]Hazard function (2) is closely analogous to the environmental damage function of Stokey (1998). While Stokey focuses on the implications of the damage function for the chosen path of x (or "z" in her notation), we will study how accelerations to the path of A affect the probability of a binary event: the occurrence of an anthropogenic existential catastrophe at any time.

[2]Ord (2020, p. 313) estimates that, as of 2020, approximately \$100 million per year was spent specifically on reducing existential risk. This is likely to be a considerable underestimate of existential safety expenditures in the sense relevant here, for two reasons. First, explicit expenditures do not include foregone consumption due to regulations that slow the development or deployment of risky technologies. Second, many efforts in e.g. nuclear non-proliferation, climate change mitigation, biosecurity, and AI safety are motivated by the desire to reduce existential risks alongside the desire to reduce damages at a smaller scale. By contrast, Moynihan (2020) argues that the very concept of an anthropogenic existential catastrophe essentially did not exist 300 years ago. To the best of our understanding, there were at that time no efforts at all taken with a view to preventing one.

model if $\beta \leq \alpha$, then an existential catastrophe is unavoidable except through indefinite degrowth, with consumption falling to zero. This immiseration would amount to the destruction of advanced civilization by other means. In the $\beta \leq \alpha$ scenario, therefore, speeding or slowing growth can have no impact on the probability of an existential catastrophe broadly construed.

Third, fixing $A_t > 0$, δ_t is assumed to be strictly convex in x_t. This imposes $\beta > 1$. The convexity implies diminishing returns to existential risk mitigation efforts.

The implications of generalizing the hazard function are discussed in Sections 5 and 6.

The probability that civilization survives to date t (starting from date 0) is given by

$$S_t \equiv e^{-\int_0^t \delta_s ds},$$

so that it corresponds to the laws of motion

$$\dot{S}_t = -\delta_t S_t, \quad S_0 = 1.$$

The probability that human civilization does *not* succumb to an anthropogenic existential catastrophe and, at least in expectation, enjoys a long and flourishing future[3] is

$$S_\infty \equiv \lim_{t \to \infty} S_t = e^{-\int_0^\infty \delta_s ds}. \tag{3}$$

Note that $S_\infty > 0$ iff $\int_0^\infty \delta_s ds$ is bounded.

2.3 Preferences

The population is fixed. The expected utility of a representative agent is

$$\int_0^\infty e^{-\rho t} S_t \, u(C_t) \, dt, \tag{4}$$

[3]In the face of natural existential risk, this will entail eventually succumbing to a natural existential catastrophe instead. From very-long-run historical data on large-scale natural catastrophes, and the typical survival rate of other mammalian species, Snyder-Beattie et al. (2019) estimate that humanity's natural existential hazard rate is "almost guaranteed to be less than one in 14,000" and "likely below one in 870,000" per year. Throughout this paper we ignore the possibility that technological advances may mitigate natural existential risks, but of course accounting for this possibility would only strengthen the headline results.

$$u(C_t) = \frac{C_t^{1-\gamma} - 1}{1-\gamma}, \quad \gamma > 1.$$

That is, flow utility $u(\cdot)$ is CRRA in consumption for some coefficient of relative risk aversion $\gamma > 1$. Flow utility is discounted at exponential rate $\rho > 0$, representing the sum of some rate of pure time preference, if any, and some rate of natural and unavoidable existential risk.

The utility of death is implicitly normalized to 0 and the death-equivalent consumption level to 1. Equivalently, we are normalizing to 1 the technology level at which, when consumption is maximized, flow utility equals 0.

A planner chooses the path of x to maximize (4) subject to (1)–(2).

Like Martin and Pindyck (2015, 2021), we impose the assumption that $\gamma > 1$ throughout the paper (except in Section 3.4). We do this in part because this appears to the empirically relevant case, as documented by Hall (1988), Lucas (1994), Chetty (2006), and others. More importantly, however, we focus on the $\gamma > 1$ case because the results are otherwise relatively uninteresting. This is for two reasons.

First, observe that when $\gamma > 1$, flow utility is upper-bounded by $\frac{1}{\gamma-1} > 0$. Accelerating consumption growth, from a baseline of positive consumption growth, therefore yields a stream of utility benefits that eventually shrinks over time. This dynamic produces the tradeoff that motivates the paper: concern for the future may cast doubt on the value of speeding technological development, because the consumption benefits of doing so primarily accrue in the short run, whereas the costs of an existential catastrophe are everlasting. By contrast, when $\gamma \leq 1$, flow utility grows in consumption without bound, so accelerations to consumption growth and reductions in existential risk can have comparable long-run benefits.

Second and relatedly, when $\gamma \leq 1$, the marginal utility of consumption does not decline quickly enough (relative to the rising value of civilization) to motivate rapid increases in consumption sacrifices for the sake of safety. As a result, the probability of long-term survival is always zero on the planner's chosen path, and accelerations or decelerations to technological development have no impact on the probability. This is detailed in Section 3.4.

3 The existential risk Kuznets curve

3.1 Optimality

Summarizing the environment of Section 2, the planner's problem is to choose $\{x_t\}_{t=0}^{\infty}$ to maximize

$$\int_0^{\infty} e^{-\rho t} S_t\, u(C_t)\, dt, \tag{5}$$

$$u(C_t) \equiv \frac{C_t^{1-\gamma} - 1}{1-\gamma}, \ \ \gamma > 1 \tag{6}$$

subject to

$$\begin{aligned}
A_0 &> 0, \\
\dot{A}_t &= gA_t \ \ (g > 0), \\
C_t &= A_t x_t, \\
S_0 &= 1, \\
\dot{S}_t &= -\delta_t S_t, \\
\delta_t &= \bar{\delta} A_t^{\alpha} x_t^{\beta} \ \ (\bar{\delta} > 0, \ \beta > \alpha > 0, \ \beta > 1).
\end{aligned} \tag{7}$$

This section finds the path of the hazard rate in the planner-optimal solution, observing that it rises and then falls with time. In the subsequent section we will explore what this implies for the impact of speeding growth on the probability of an anthropogenic existential catastrophe. From now on, we will typically refer to such an event simply as a "catastrophe".

The planner faces one choice variable, x_t, and one state variable, S_t. Her (expected) flow payoff at t is $S_t u(C_t)$. Her problem can be represented by the following current-value Hamiltonian:

$$\begin{aligned}
\mathcal{H}_t &= S_t u(C_t) + v_t \dot{S}_t \\
&= S_t \frac{(A_t x_t)^{1-\gamma} - 1}{1-\gamma} - v_t\, \bar{\delta} A_t^{\alpha} x_t^{\beta} S_t,
\end{aligned} \tag{8}$$

where

$$v_t = \int_t^{\infty} e^{-\rho(s-t)} \frac{S_s}{S_t} u(C_s) ds \tag{9}$$

is the costate variable on survival: the expected value of the future of civilization at t, conditional on survival to t.[4]

On an optimal path, the first-order condition on (8) with respect to the choice variable x_t is satisfied. Differentiating (8) with respect to x_t, we have

$$S_t A_t^{1-\gamma} x_t^{-\gamma} - \bar{\delta} A_t^{\alpha} \beta x_t^{\beta-1} v_t S_t \geq 0, \tag{10}$$

with inequality iff the left-hand side is positive at $x_t = 1$, in which case $x_t = 1$ is optimal.[5] Thus,

- As long as (10) is nonnegative at $x_t = 1$, the optimal choice of $x_t \in [0, 1]$ equals 1. Even the first marginal sacrifices of consumption would come with greater flow costs than expected benefits.

- When (10) is negative at $x_t = 1$, the optimal choice of x_t is interior. It sets (10) equal to zero, maintaining the condition that the marginal cost to flow utility of lowering consumption equals the expected benefit via risk reduction.[6]

In fact there is a unique[7] optimal path, characterized by first-order condition (10), a first-order condition corresponding to the state variable S_t, and identity (9). This is shown in Appendix A.1 for the strictly more general environments of Sections 5 and 6. Throughout this section, however, our discussion will rely only on the observation that on any optimal path, (10) must be satisfied and on any feasible path, v_t is upper-bounded by

$$\bar{v} \equiv \frac{1}{\rho(\gamma - 1)}. \tag{11}$$

3.2 Initial risk increases

The condition that (10) is nonnegative at $x_t = 1$ is equivalent to the condition that

$$A_t^{-(\alpha+\gamma-1)} \geq \bar{\delta} \beta v_t. \tag{12}$$

[4]The fact that the costate variable on survival must equal (9) can be seen immediately by reflecting on the fact that, in effect, the value of saving the world must equal the value of the world; but it is also derived formally in Appendix A.1.

[5]The second derivative with respect to x_t is negative by the assumption that $\beta > 1$.

[6]We can ignore the possibility that the optimal choice of x_t equals 0 because such a choice yields infinite flow disutility.

[7]Under the restriction of piecewise continuity. If x is optimal, measure-zero deviations from x are of course also optimal.

The continuation value of civilization at t given survival to t, v_t, always strictly rises over time. This follows from the fact that, given the optimal paths $\{C_s\}_{s\geq t}$ and $\{\delta_s\}_{s\geq t}$ achievable at a given initial technology level A_t, a higher initial technology level allows for a path with an equal hazard rate but more consumption at each future period, by the assumption that $\beta > \alpha$. A higher initial technology level always enables the planner to implement a preferred future.

Therefore, early in time, when A_t is low, inequality (12) is satisfied. The optimal policy choice is $x = 1$, and the hazard rate rises with A at rate αg. We will assume that time 0 is defined to be early enough in time that inequality (12) is satisfied strictly at $t = 0$.

3.3 Eventual risk declines and survival

As the left-hand side of (12) falls exponentially with A_t and the right-hand side rises, there is a unique time t^* at which (12) holds with equality. After t^*, the optimal choice of x_t is interior and sets (10) equal to zero.

Setting (10) equal to zero, rearranging, and taking the growth rate of each side, we can find the growth rate of the policy choice variable:

$$x_t^{1-\beta-\gamma} = \bar{\delta}\beta A_t^{\alpha+\gamma-1} v_t \tag{13}$$

$$\implies g_{xt} = -\frac{\alpha+\gamma-1}{\beta+\gamma-1}g - \frac{1}{\beta+\gamma-1}g_{vt}, \tag{14}$$

where, given a time-dependent variable y, $g_{yt} \equiv \dot{y}_t/y_t$ denotes its proportional growth rate at t.

The hazard rate in turn grows as

$$\begin{aligned} g_{\delta t} &= \alpha g + \beta g_{xt} \\ &= -\frac{(\beta-\alpha)(\gamma-1)}{\beta+\gamma-1}g - \frac{\beta}{\beta+\gamma-1}g_{vt}. \end{aligned} \tag{15}$$

Because $\beta > \alpha$ and $\gamma > 1$, (15) is negative.

Furthermore, though g_{vt} is always positive, $g_{vt} \to 0$. This roughly follows from the fact that the expected value of the future v_t is bounded above by $\bar{v}$.[8] This gives us the asymptotic long-run negative growth rates g_x and g_δ.

[8] The $g_{vt} \to 0$ limit is shown formally in Appendix A.2.

Finally, since $C_t = A_t x_t$, we have

$$\begin{aligned} g_{Ct} &= g + g_{xt} \\ &= \frac{\beta - \alpha}{\beta + \gamma - 1} g - \frac{1}{\beta + \gamma - 1} g_{vt}. \end{aligned}$$

Because $\beta > \alpha$, long-run consumption growth is positive: though x declines to 0, A grows more quickly than x declines. Indeed, the growth of consumption is key to the growth in sacrifices for the sake of safety. In the face of decreasing marginal utility to *consumption* and decreasing marginal returns to *safety effort*, potential consumption increases are split between the former and latter so that the marginal value of each remains equal.

To summarize:

Proposition 1. ***The existential risk Kuznets curve***
On the planner-optimal path defined by (5)–(7), there exists a time t^* such that for $t \leq t^*$,

$$\begin{aligned} x_t &= 1, \\ g_{Ct} &= g > 0, \\ g_{\delta t} &= \alpha g > 0 \end{aligned}$$

and for $t > t^*$,

$$\lim_{t\to\infty} g_{xt} = -\frac{\alpha + \gamma - 1}{\beta + \gamma - 1} g < 0, \tag{16}$$

$$\lim_{t\to\infty} g_{Ct} = \frac{\beta - \alpha}{\beta + \gamma - 1} g > 0,$$

$$\lim_{t\to\infty} g_{\delta t} = -\frac{(\beta - \alpha)(\gamma - 1)}{\beta + \gamma - 1} g < 0 \tag{17}$$

with all three limits approached from below.

Corollary 1.1. ***Survival***
On the planner-optimal path defined by (5)–(7), $S_\infty > 0$.

Proof. The result follows from (17) and the definition of S_∞. Because δ_t ultimately falls exponentially, $\int_0^\infty \delta_t dt < \infty$, so $S_\infty \equiv e^{-\int_0^\infty \delta_t dt} > 0$.

Note that $\delta_t \to 0$ is insufficient for survival. If δ_t fell to 0 too slowly, the integral would diverge, and we would have $S_\infty = 0$. □

3.4 No survival with $\gamma \leq 1$

As noted in Section 2.3, one reason for focusing on the $\gamma > 1$ case is that, when the marginal utility of consumption declines too slowly, a rapid shift from consumption to safety effort is not implemented, and the probability of long-term survival is always zero.

Proposition 2. ***Policy choice and risk with*** $\gamma \leq 1$
Suppose a planner faces problem (5)–(7), but with utility function (6) replaced by

$$u(C_t) = \begin{cases} \log(C_t), & \gamma = 1; \\ \frac{C_t^{1-\gamma}-1}{1-\gamma}, & \gamma < 1 \end{cases} \tag{18}$$

for some $\gamma \leq 1$*, and*

$$\rho > \underline{\rho} \equiv \frac{(\beta - \alpha)(1-\gamma)}{\beta} g \tag{19}$$

to ensure the existence of an optimal policy.

Then there exists a time t^* *such that for* $t \leq t^*$,

$$\begin{aligned} x_t &= 1, \\ g_{Ct} &= g > 0, \\ g_{\delta t} &= \alpha g > 0 \end{aligned}$$

and for $t > t^*$,

$$\lim_{t \to \infty} g_{xt} = -\frac{\alpha}{\beta} g < 0, \tag{20}$$

$$\lim_{t \to \infty} g_{Ct} = \frac{\beta - \alpha}{\beta} g > 0,$$

$$\lim_{t \to \infty} \delta_t t = \frac{\rho}{(\beta - \alpha) g} > 0, \quad \gamma = 1; \tag{21}$$

$$\delta^* \equiv \lim_{t \to \infty} \delta_t = \frac{(\rho - \underline{\rho})(1-\gamma)}{\beta + \gamma - 1} > 0, \quad \gamma < 1. \tag{22}$$

Proof. See Appendix A.2. □

Corollary 2.1. ***No survival with*** $\gamma \leq 1$
On the planner-optimal path defined by (5)–(7), with utility function (6) replaced by (18), $S_\infty = 0$*.*

Proof. The result follows from (21)–(22) and the definition of S_∞. When δ_t is asymptotically constant or proportional to $1/t$, $\int_0^\infty \delta_t dt = \infty$, so $S_\infty \equiv e^{-\int_0^\infty \delta_t dt} = 0$. □

The case in which δ_t declines proportionally to $1/t$, obtained by $\gamma = 1$, is the edge case in which the expected length of time until a catastrophe is infinite even though the probability of catastrophe is 1.

Though a catastrophe is here inevitable on the chosen path, it can be seen from (22) that faster technology growth g lowers the asymptotic hazard rate δ^* when $\gamma < 1$. This is essentially because, when $\gamma < 1$, consumption and thus flow utility grow at a higher exponential rate in the long run when g is higher, so the effect of raising g is similar to the effect of decreasing the discount rate ρ.

There is not a general result that increases to g always increase the "life expectancy of civilization" when $\gamma < 1$, however. This is discussed briefly at the end of Section 4.2.

Understanding the path of policy choice and risk is somewhat more complex when $\gamma \leq 1$ than when $\gamma > 1$, because we do not have the result that v_t is asymptotically constant, but a sketch is as follows.

As in the $\gamma > 1$ setting of Proposition 1, early in time inequality (12) holds and it is optimal to set $x_t = 1$. Likewise, later in time, optimality requires setting $x_t < 1$ so as to maintain

$$\begin{aligned} A_t u'(C_t) &= \frac{\partial \delta}{\partial x} \cdot v_t \\ \implies A_t x_t \, C_t^{-\gamma} &= \bar{\delta} A_t^\alpha \beta x_t^\beta v_t \\ \implies \delta_t &= \frac{C_t^{1-\gamma}}{\beta v_t}. \end{aligned} \tag{23}$$

Observe from (9) that v_t grows roughly with flow utility $u(C_t)$. Flow utility, for large C_t, then grows approximately like $C_t^{1-\gamma}$ when $\gamma < 1$. The result is that, though consumption grows exponentially in the long run for any value of γ, δ is asymptotically constant when $\gamma < 1$.

Intuitively, for the policy path to be optimal, it must maintain

a) the flow utility to proportionally increasing consumption, $C_t \cdot C_t^{-\gamma}$

$=$

b) the damage done via proportionally raising the hazard rate, which equals <u>the hazard rate</u> × <u>the value of civilization</u>.

When the value of civilization also grows like $C_t^{1-\gamma}$, as it does when $\gamma < 1$, the hazard rate must be constant for (a) and (b) to grow at the same rate. When $\gamma > 1$, the value of civilization is asymptotically constant, so the hazard rate falls like $C_t^{1-\gamma}$.

When $\gamma = 1$, given that consumption grows exponentially, $\log(C_t)$ and thus v_t grow linearly. The hazard rate then falls proportionally to $1/t$.

To focus on the scenarios in which accelerations to technological development can affect the probability of survival, and for simplicity, we will maintain the $\gamma > 1$ assumption throughout the remainder of the paper.

3.5 Simulation

The paths of policy choice and the hazard rate are simulated below, for the following parameter values:

ρ	0.02	$\bar{\delta}$	0.00012
γ	1.5	α	1
g	0.02	β	2
A_0	2		

Table 1: Simulation parameters for Figure 1

The values of ρ, γ, 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 at $t = 75$ is four times consumption per capita, roughly matching estimates from Klenow et al. (2023).[9] That is, the first year of the simulation might be taken to denote 1949, the year at which a nuclear war between superpowers first became possible, and the 75th year might be taken to denote the time of writing. $\bar{\delta}$, α, and β are chosen so that the hazard rate today

[9]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, including many which are poorer than the United States.

is approximately 0.1%, matching Stern's (2007) oft-cited figure; so that the hazard rate begins to fall at approximately $t = 100$; and so that the growth rate and then the decay rate of the hazard rate are non-negligible, for clarity in illustration.

The probability of survival S_∞ under these parameters, from $t = 75$ onward, is approximately 65%.

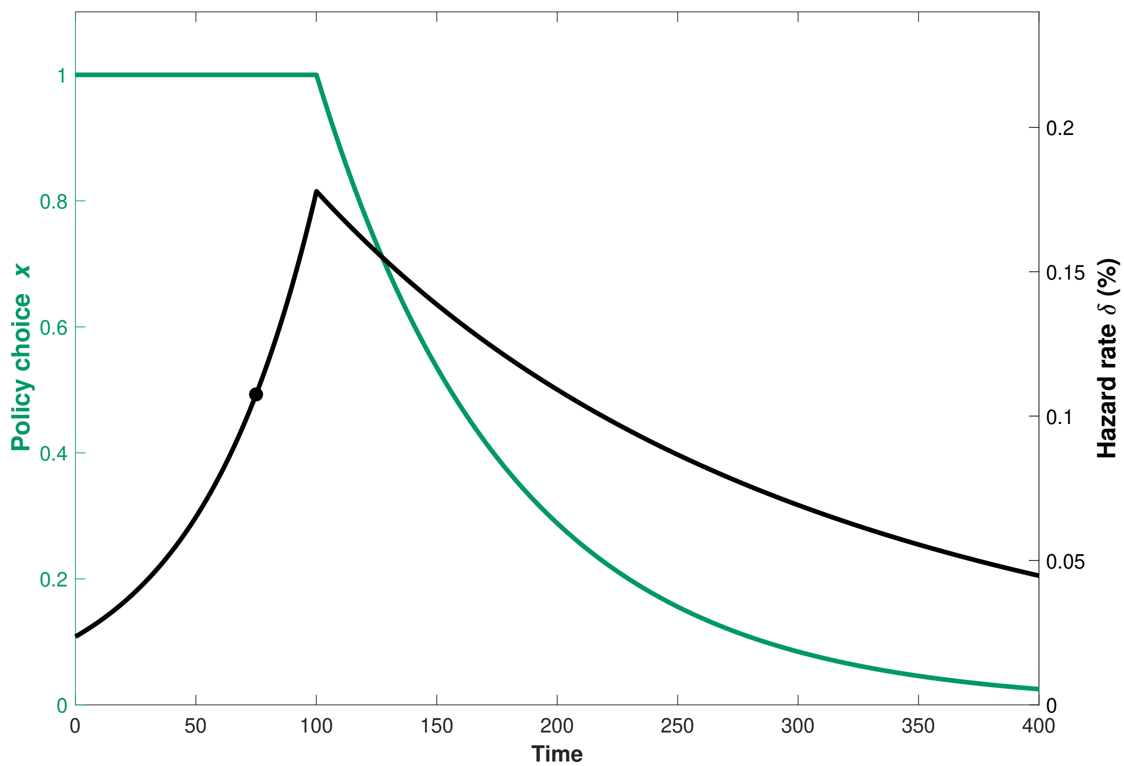


Figure 1: Evolution of the policy choice and the hazard rate along the optimal path

Calculations and code for replicating the simulation and corresponding probability of survival may be found in Appendix B.

As Figure 1 illustrates, one potentially unappealing feature of the baseline model is that it implies that, on the optimal path, the hazard rate only rises during the regime in which no sacrifices whatsoever are made for existential safety. In this respect it closely resembles Stokey's (1998) "environmental Kuznets curve", which also features damages which rise exponentially with economic growth and then fall sharply past the point at which it first becomes optimal to take action. As discussed in Section 2.2, this pattern may be at odds with the experience of the last century, during which the hazard rate has arguably risen alongside existential risk mitigation efforts.

As in Stokey (1998), this dynamic is essentially driven by the lack of a lower Inada condition on $1-x$. If marginal "safety expenditures" lower the hazard rate infinitely per unit spent at $x=1$, then as long as $v_t > 0$ it is optimal to set $x_t < 1$, even if early in time the hazard rate is allowed to rise. Rising δ can thus be found alongside falling x by tweaking the behavior of the hazard function around $x=1$. Such tweaks do not affect the long-run behavior of the policy choice or the hazard rate as given by (16)–(17), which are determined by the behavior of the hazard function around $x=0$. This is discussed further in Section 5.5.

4 How does speeding growth affect risk?

The analysis of the previous section lets us determine how various shocks to the technology growth path affect the probability of survival in the planner's solution, S_∞. As we will see, while the impact on risk of a temporary shock is ambiguous, the impact of a permanent level or growth effect is always to lower risk. Thus, while the possibility of an existential risk Kuznets curve supports the contention that existential risk reduction is overwhelmingly valuable from a low-discount-rate perspective, this possibility generates a case that speeding growth is beneficial from such a perspective. In the face of an existential risk Kuznets curve, the appearance of a tradeoff between existential risk and growth may be only a short-term illusion.

The impact of a shock to growth on the probability of survival is explored for a more general class of hazard functions in Section 5.4, and the results are stated formally there in Proposition 6. Here we will illustrate the dynamics with more discussion using hazard function (2) in particular.

4.1 Change of variables

Absent recession or long-term stagnation, A crosses every value from A_0 to ∞ exactly once. So the area under the hazard curve can be defined by integrating with respect to A instead of t. We will refer to the area under the hazard curve as "cumulative risk", denoted X, and define it as a function of the technology path $A(\cdot)$, assuming that the policy path is the optimal path given $A(\cdot)$, denoted $x[A(\cdot)]$.

$$X\big(A(\cdot)\big) \equiv \int_0^\infty \bar{\delta} A_t^\alpha x_t[A(\cdot)]^\beta dt = \int_{A_0}^\infty \bar{\delta}\, A^\alpha\, x_A[A(\cdot)]^\beta\, dA \left(\frac{dA}{dt}\right)^{-1}$$
$$= \int_{A_0}^\infty \bar{\delta}\, A^\alpha \dot{A}_A^{-1}\, x_A[A(\cdot)]^\beta\, dA, \tag{24}$$

where x_A and $\dot{A}_A$ denote the values of x and $\dot{A}$ when the technology level equals the subscripted A. Expression (24) for the area under the curve will make it easier to see how shocks to growth affect cumulative risk.

Observe from (3) that the probability of survival is monotonically decreasing in cumulative risk, with $S_\infty = e^{-X}$.

4.2 Three ways of speeding growth

Temporary level effects

The effect of a temporary positive shock to the technology level A_t, letting x_t adjust instantaneously, depends on whether the shock occurs before or after the regime-change time t^*.

Before t^*, temporarily raising A has no impact on the optimal choice of x.[10] δ thus rises. The future hazard rate is unaffected, so cumulative risk increases.

After t^*, temporarily multiplying A_t by $m > 1$ multiplies the optimal choice of x_t by $m^{-\frac{\alpha+\gamma-1}{\beta+\gamma-1}}$, by (13). In combination, the positive shock to A_t and the negative impact on x_t multiply δ_t by $m^{\alpha-\beta\frac{\alpha+\gamma-1}{\beta+\gamma-1}} = m^{-\frac{(\beta-\alpha)(\gamma-1)}{\beta+\gamma-1}} < 1$. This decreases cumulative risk.

[10] Unless the increase to A is large enough to reverse inequality (12), a case we will ignore for simplicity.

Permanent level effects

Consider the effect of multiplying A_t by $m > 1$, and subsequently maintaining exponential growth in A. Observe that the initial value of A_t alone determines the optimal subsequent path of x_t. The level effect in A therefore amounts to a "leap forward in time": a slice cut out of the existential risk Kuznets curve. Cumulative risk falls from (24) to

$$\int_{A_0}^{A_t} \bar{\delta}\, A^{\alpha} \dot{A}_A^{-1}\, x_A^{\beta}\, dA + \int_{mA_t}^{\infty} \bar{\delta}\, A^{\alpha} \dot{A}_A^{-1}\, x_A^{\beta}\, dA.$$

More generally, we might model a level effect as an increase to $\dot{A}$ at some range of values of A (say, from A_t to mA_t for $m > 1$). Because the exponent on $\dot{A}$ in the integral is negative, acceleration lowers the risk endured at the given range of technology levels. A discontinuous jump in the technology level amounts to raising $\dot{A}_A$ to ∞, and thus lowering $\dot{A}_A^{-1}$ to 0, from $A = A_t$ to mA_t.

In either case, such a leap may temporarily increase the hazard rate, if it begins on the increasing side of the curve, so it may appear to contemporaries to be increasing the risk of a catastrophe. However, a level effect (with immediate adjustment in policy choice) actually decreases cumulative risk.

The effects of a sharp and permanent level effect are illustrated in Figure 2. The parameter values used to illustrate the baseline path are the same as those used to simulate Figure 1. The level effect takes place "today", at $t = 75$, and multiplies A by $e^{0.2}$, so that at $g = 0.02$, it amounts to a 10-year leap forward.

Recall from Section 3.5 that the probability of survival (from $t = 75$ onward) on the baseline path is approximately 65%. The proportional increase in the probability of survival can be found analytically. Cumulative risk X declines by precisely the area under the baseline hazard curve from $t = 75$ to 85; and since $\delta_{75} = 0.1\%$, $g = 0.02$, and $\alpha = 1$, this difference equals

$$\Delta X = -0.001 \int_0^{10} e^{0.02t} dt = -0.05(e^{0.2} - 1).$$

$S_\infty = e^{-X}$ is then multiplied by

$$e^{-\Delta X} \approx 1.011,$$

so that in absolute terms S_∞ rises by approximately $0.65 \cdot 0.011 \approx 0.7\%$.

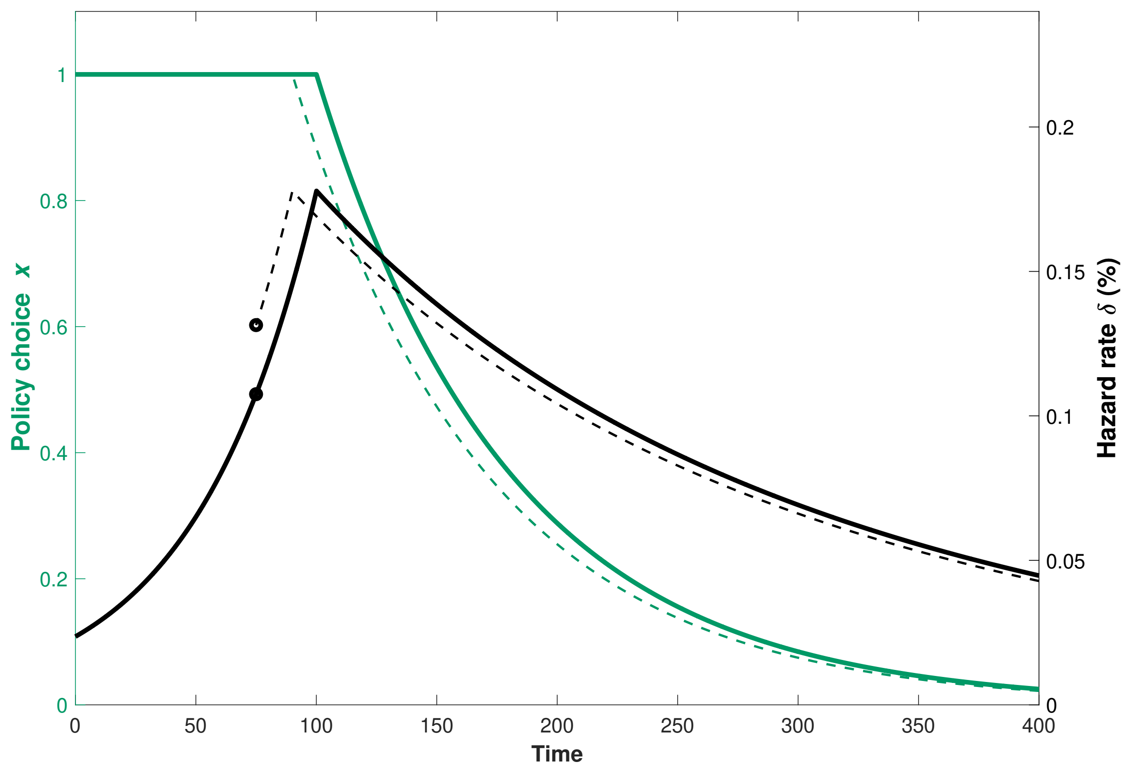


Figure 2: A level effect to growth shrinks cumulative risk

Calculations and code for replicating the simulation may be found in Appendix B.

To state this lesson in reverse, consider the implications of a large negative level effect today, which returned the world to a state of ignorance about every technology developed since 1924. We would largely be doomed to relive the nuclear standoffs, emissions-intensive industrializations, and biotechnological hazards of the past. With enough resets of this kind, a catastrophe would presumably be inevitable.

Growth effects

Growth effects shrink cumulative risk for two reasons.

First, at every moment t after the time τ initiating the growth effect, A_t is higher than it would have been. In particular, if at time τ the growth rate is multiplied by $m > 1$, then A_t (for $t > \tau$) takes the value that $A_{\tau+m(t-\tau)}$ would have taken in the absence of the growth effect. Even if x_t (for $t > \tau$) only adjusts to the value that would be optimal if the growth effect had no effect on v_t, therefore, then a growth effect effectively rescales the hazard curve after τ, dividing its area by m.

Second, though typically less importantly, growth effects increase the value of the future v_t at each $t > \tau$. By (12), this motivates lower choices of x_t.

Unlike level effects, growth effects can even render survival possible when it would otherwise have been impossible. In the baseline model introduced in Section 2 and solved in Section 3, a positive probability of survival is always feasible and always chosen on the planner-optimal path. Consider the implications of full stagnation: a negative growth effect permanently setting $g = 0$. The hazard rate is then permanently positive, and so survival is impossible, though it would have been possible at any positive technology growth rate.

More subtly, consider a negative growth effect in which the technology level A_t grows not exponentially at rate g but power-functionally, so that $A_t = t^k$ for some $k > 0$. The exponential growth rate of A is then not constant at g but time-varying, with $g_{At} = k/t$. By (15) (and recalling that g_v is asymptotically zero), it then follows that δ_t asymptotically falls to zero like $t^{-\frac{(\alpha-\beta)(\gamma-1)}{\beta+\gamma-1}k}$. Since cumulative risk is finite for $\delta_t \propto t^{-\kappa}$ iff $\kappa > 1$, the probability of survival is positive given $A_t = t^k$ iff $k > \frac{\beta+\gamma-1}{(\alpha-\beta)(\gamma-1)}$.

Abstracting from the details of the model at hand, the more general lesson is hopefully clear. Adding or removing a finite slice from a finite or infinite area leaves it finite or infinite respectively, but foreshortening a heavy-tailed curve with an infinite integral can yield a thin-tailed curve whose integral is finite.

Importantly, the result that stagnation is deadly is driven by the assumption that the hazard rate is always greater than zero. Given this assumption, a positive probability of long-term survival can only be achieved by quickly driving the hazard rate toward zero, a process which presumably requires technological innovation. Models like that of Baranzini and Bourguinion (1995), or the Jones (2016) “Russian roulette” model, produce the result that technological stagnation yields safety because they assume that, given stagnation, the hazard rate is zero, at least if it occurs at some potential levels of consumption or technological development. The implications of this alternative assumption are explored further in Section 6.

Also, in the $\gamma < 1$ case of Section 3.4 in which catastrophe is inevitable, positive growth effects do not necessarily increase "civilizational life expectancy". In that setting, stagnation at a low level of technological development A yields a permanent hazard rate of $\bar{\delta}A^{\alpha}$, which may be arbitrarily low, and thus an expected duration until catastrophe of $1/(\bar{\delta}A^{\alpha})$, which may be arbitrarily high. Raising g to a large positive number can then (perhaps quickly) yield hazard rates that permanently approximate δ^*, lowering civilizational life expectancy to approximately $1/\delta^*$.

4.3 Patience vs. growth

The key mechanism at work in this paper is that as consumption grows, people's willingness to sacrifice consumption for safety rises. By contrast, those concerned about the security of the long-term future often prioritize moral persuasion, appealing to ethical arguments for a low rate of pure time preference. Consider e.g. the Stern–Nordhaus debate (and the long debate since) over the social discount rate to use in the context of climate policy, or the arguments for concern for the future put forward by philosophers such as Parfit (1984), Cowen and Parfit (1992), and more recently Ord (2020) and MacAskill (2022). How do the these two mechanisms—a permanent level effect to A vs. a permanent reduction to the rate of pure time preference ρ—compare in terms of increasing the probability of survival?

We will see that, early in time, decreases to ρ are arbitrarily more impactful than increases to A. Late in time, however, the impacts of the two interventions are comparable.

A permanent level effect at t, whereby A is multiplied by m slightly greater than 1, amounts to a leap forward in time of approximately m/g years. This decreases cumulative risk by approximately $\delta_t m/g$.

Before the regime-change time t^*, therefore, the impact of a level effect on cumulative risk rises exponentially with δ_t over time. Early in time, when A_t and δ_t are arbitrarily low, the impact of the level effect on cumulative risk is arbitrarily low. The impact of a decrease to ρ on cumulative risk, on the other hand, does not change over time before t^*. A decrease to ρ does not affect the hazard rate immediately, but decreases it in the future by pulling forward the regime-change time and changing the path of the hazard rate afterward. These impacts do not depend on *when* (before t^*) ρ is lowered.

By contrast, consider what happens as $u(c_t) \to \frac{1}{\gamma-1}$ and thus $v_t \to \frac{1}{\rho(\gamma-1)}$. By (13), in the limit,

$$x_t \approx \left(\frac{\bar{\delta}\beta}{\rho(\gamma-1)}\right)^{-\frac{1}{\beta+\gamma-1}} A_t^{-\frac{\alpha+\gamma-1}{\beta+\gamma-1}}. \tag{25}$$

At large t, permanently multiplying A by $m > 1$ multiplies x_s, at each subsequent period $s \geq t$, by approximately $m^{-\frac{\alpha+\gamma-1}{\beta+\gamma-1}}$. In conjunction, the increase to A_s and the proportional decrease to x_s multiply δ_s by $m^{-\frac{(\beta-\alpha)(\gamma-1)}{\beta+\gamma-1}}$ for $s \geq t$. Similarly, permanently dividing ρ by $m > 1$ multiplies x_s ($s \geq t$) by approximately $m^{-\frac{1}{\beta+\gamma-1}}$, which multiplies δ_s ($s \geq t$) by approximately $m^{-\frac{\beta}{\beta+\gamma-1}}$. The impacts are equal iff

$$\begin{aligned} (\beta-\alpha)(\gamma-1) &= \beta \\ \iff \gamma &= 2 + \frac{\alpha}{\beta-\alpha}, \end{aligned} \tag{26}$$

with the level effect more impactful if the left-hand side is greater and the decrease to ρ more impactful if the right-hand side is greater. The growth-based intervention is more impactful when γ is higher, because higher values of γ motivate faster transitions from consumption to risk-reduction.

Since $\beta > \alpha > 0$, expression (26) reveals that the level effect can only be more impactful in this model if $\gamma > 2$. Still, it is notable that mere level effects to growth can ultimately affect the probability of survival at a comparable scale to permanent, equally-proportioned decreases to the social rate of pure time preference (holding technology growth fixed). Put another way, even temporary stagnation can carry long-term costs similar to those of permanently moving ethical attitudes away from concern for the future.

5 Generalization

The results of Sections 3 and 4 are set in the economic environment of Section 2. The three central ingredients of this environment are of course the growth path of technology, the hazard rate as a function of technology and policy, and the utility function. A particular functional form is assumed for each.

Throughout this section we will maintain the assumption of a CRRA utility function with $\gamma > 1$. We will however greatly relax our assumptions on the technology path and the hazard rate.

5.1 Assumptions

Assumptions on technology growth

Instead of assuming that technology grows exponentially, we will assume only that $A(t)$ satisfies the following conditions:

A1. continuous differentiability,

A2. strict monotonicity,

A3. $\lim_{t\to-\infty} A(t) = 0$, and

A4. $\lim_{t\to\infty} A(t) = \infty$.

We will call a technology path $A(\cdot)$ admissible if it satisfies A1–A4.

We will continue to treat the growth of technology as exogenous. Importantly, this is without loss of generality for our purposes. We are concerned with the implications of a given technology path for optimal policy and cumulative risk, not with how a given technology path is generated. We therefore do not need to model how the rate of technological development may itself be determined by a planner's funding of research and development, innovation by market participants, or any other forces.

Assumptions on the hazard rate

We will also consider a wider class of hazard functions. Among these, we will find relatively simple conditions under which a given hazard function and a given technology growth path are compatible with survival on the planner-optimal policy. In Sections 5.2–5.3, generalizing Proposition 1, we find that growth motivates increasing concern for safety: it is often optimal to set $x = 1$ early in time and $x \to 0$ late in time. Except in cases where lowering risk is so difficult that it is not achieved even with stagnation in consumption, the hazard rate is also driven to 0. In Section 5.4, generalizing Section 4, we likewise find that when a hazard function is compatible with survival, faster growth in consumption technology generally increases the probability of survival. The results support the robustness of the primary lessons drawn from hazard function (2): that survival is likely possible on the optimal path, and that faster consumption technology growth, if efficiently regulated, will make raise its probability.

In Section 5.5, we will identify two particular hazard functions in this class that may be of interest. The first illustrates that, early in time, the hazard rate may increase alongside smooth declines in x. The second is "microfounded" by an assumption that increases in safety expenditure lower risk through redundant safeguards.

Return to the three desiderata preceding the introduction of hazard function (2). We will assume weakenings of two of these desiderata directly, and certain results will require a weakening of the third. In particular, we will universally assume that the hazard rate increases in x no less quickly than in A and is weakly convex in x. For certain results we will assume that the hazard rate does not decrease too quickly in A.

We will add to these the preliminary conditions that $\delta(\cdot)$ is continuously differentiable; that, when consumption equals zero, so that the entire productive capacity of society is dedicated to existential risk reduction, $\delta = 0$; and that otherwise $\delta > 0$.[11]

Formally, we will assume at most that the hazard rate is a function of $A > 0$ and $x \in (0, 1]$ satisfying the following conditions:

- D1. $\delta(A, x) > 0$,
- D2. $\lim_{x\to 0} \delta(A, x) = \lim_{A\to 0} \delta(A, x) = 0$,
- D3. twice continuous differentiability,[12]
- D4. $\eta_x(A, x) \geq \eta_A(A, x)$, and
- D5. weak concavity in x,

where η_y denotes the elasticity of δ with respect to $y \in \{A, x\}$. We will call a hazard function admissible if it satisfies D1–D5.

[11] Recall that the hazard rate denotes the flow probability of *anthropogenic* existential catastrophe.

[12] We will define $\frac{\partial \delta}{\partial y}(A, 1) \equiv \lim_{x\to 1} \frac{\partial \delta}{\partial y}(A, x)$ for $y \in \{A, x\}$, and allow these derivatives to be infinite.

Note that the constant elasticity hazard function of Sections 2–4 is admissible, with $\eta_A = \alpha$ and $\eta_x = \beta$ independent of A and x. Note also that we do not require $\eta_A(A, x)$ always to be positive. That is, we will allow for the possibility that new technologies sometimes lower the hazard rate at a given degree of foregone consumption.

5.2 Stagnation vs. unbounded consumption

Let $C^* \equiv \lim_{t\to\infty} A_t x_t$, when this limit is defined.

Given hazard function (2), $C^* = \infty$. This follows from Proposition 1. Since $\lim_{t\to\infty} g_{xt} = -\frac{\alpha+\gamma-1}{\beta+\gamma-1} g$, $\lim_{t\to\infty} g_{Ax,t} = (1 - \frac{\alpha+\gamma-1}{\beta+\gamma-1}) g$, which is positive by the assumption that $\beta > \alpha$. In the long run, consumption growth is positive and exponential.

However, some hazard functions satisfying D1–D5 motivate decreases to x fast enough that we do not have $C^* = \infty$. C^* may be finite, or C_t may oscillate indefinitely without growing ever higher.

Proposition 3. ***Stagnation vs. unbounded consumption***
Define

$$\begin{aligned} R(C) &\equiv \lim_{A\to\infty} \frac{\partial \delta}{\partial x}\left(A, \frac{C}{A}\right) \frac{C^\gamma}{A} \bar{v}, \\ R^* &\equiv \lim_{C\to\infty} R(C). \end{aligned} \tag{27}$$

Given an admissible technology path and hazard function,

a) *If $R^* \leq 1$, then $C^* = \infty$.*

b) *If $R^* > 1$, then $C^* \neq \infty$.*

Proof. See Appendix A.3. □

To interpret the result, recall that $x = C/A$. The limit in (27) characterizes, if C is fixed even as A grows, what happens to the ratio of the marginal value of lowering x via increased safety ($\frac{\partial \delta}{\partial x} \cdot v$) to the marginal utility of raising x via increased consumption ($AC^{-\gamma}$). If the ratio approaches 1, then it is optimal for consumption to stagnate in the long run at C. If the ratio is greater than 1 for sufficiently large C, therefore, then stagnation at some finite C is optimal. If the ratio is less than or equal to 1 even as $C \to \infty$, then stagnation is not optimal.

Recall from (11) that $\bar{v} \equiv \frac{1}{\rho(\gamma-1)}$. When $R(C) > 0$, therefore, $R(C)$ is decreasing in ρ. A lower discount rate ρ can thus shift R^* from below to above 1, resulting in stagnation when there would otherwise have been long-run consumption growth, but never the reverse. There is no general result that consumption stagnation is desirable when ρ is sufficiently low, or undesirable when ρ is sufficiently large: for many hazard functions, as implicitly shown at the end of Section 5.3, R^* is above 1 (even infinite) or below 1 (even 0) for any $\rho > 0$. Still, Proposition 3 illustrates how calls for an "end to growth" of some kind may be compatible with the results at the heart of this paper. Concern for the future can motivate controls on technological deployment strict enough to halt growth in *consumption*, despite the tendency for accelerating (even impatiently-regulated) technological *development* to lower cumulative risk.

5.3 The Kuznets curve generalized

Proposition 4. ***The Kuznets curve generalized***
Given an admissible technology path and hazard function,

a) $\lim_{t\to-\infty} x_t = 1$.
If η_A *is bounded above* $1-\gamma$*, then* $\lim_{t\to\infty} x_t = 0$.

b) $\lim_{t\to-\infty} \delta_t = 0$.
If $C^* = \infty$*, then* $\lim_{t\to\infty} \delta_t = 0$.
If $C^* \neq \infty$*,* η_A *is bounded above* $1-\gamma$*, and* η_x *is upper-bounded, then* $\lim_{t\to\infty} \delta_t \neq 0$.

Proof. The proof of (a) is given in Appendix A.4. The proof of (b) is as follows.

By D1, D2, and D5, $\delta(A,x)$ is non-decreasing in x. So for all t, $\delta_t \leq \delta(A_t, 1)$. By D2, $\lim_{A\to 0} \delta(A,1) = 0$. So by A3, $\lim_{t\to-\infty} \delta_t = 0$.

For the positive limit, begin with the weak first-order condition that the marginal flow utility of increasing x must weakly exceed the marginal cost via an increased hazard rate. Then multiply both sides by x_t:

$$\begin{aligned} A_t^{1-\gamma} x_t^{-\gamma} &\geq \frac{\partial\delta}{\partial x}(A_t, x_t)\, v_t \\ \implies (A_t x_t)^{1-\gamma} &\geq \frac{\partial\delta}{\partial x}(A_t, x_t) x_t\, v_t. \end{aligned} \tag{28}$$

If $C^* = \infty$, the left-hand side of (28) tends to 0. Since v is (eventually) positive and does not fall by D4, $\frac{\partial \delta}{\partial x} x \to 0$. Since $\frac{\partial \delta}{\partial x} x \geq \delta$ by D1 and D5, $\delta \to 0$.

If η_A is bounded above $1-\gamma$, $\lim_{t\to\infty} x_t = 0$ by (a). Since eventually $x_t < 1$, eventually (28) holds with equality. If $C^* \neq \infty$, the left-hand side does not tend to 0 in the limit. Because v_t is upper-bounded, $\frac{\partial \delta}{\partial x} x$ does not tend to zero either. So if $\eta_x \equiv \frac{\partial \delta}{\partial x}\frac{x}{\delta}$ is upper-bounded, $\delta \not\to 0$. □

Part (b) of the proposition stems from the fact that, as long as consumption rises without bound, its marginal utility falls to zero. If the hazard rate does not also fall to zero, the marginal value of sacrificing consumption to lower it further stays positive. The hazard rate must therefore fall to zero.

Even so, unbounded consumption growth does not necessarily coincide with a positive probability of survival. To achieve $S_\infty > 0$, δ_t must not only fall to 0 but fall sufficiently quickly. This in turn is guaranteed whenever consumption rises sufficiently quickly, which holds under a strengthening of the condition for unbounded consumption growth from Proposition 3.

Proposition 5. ***Survival generalized***
Given an admissible technology path such that, for some $k > 1$ and some $\underline{t}$ we have

$$A_t \geq t^{\frac{k}{\gamma-1}} \quad \forall t > \underline{t}, \tag{29}$$

define

$$\tilde{R}(k) \equiv \lim_{t\to\infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{k}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{k\gamma}{\gamma-1}}}{A_t} \bar{v}.$$

Given an admissible technology path satisfying (29) and an admissible hazard function,

a) *If $\lim_{k\downarrow 1} \tilde{R}(k) < 1$, then $\exists \underline{t} \,:\, C_t > t^{\frac{1}{\gamma-1}} \;\; \forall t > \underline{t}$ and $S_\infty > 0$.*

b) *If $\lim_{k\uparrow 1} \tilde{R}(k) > 1$, then $\exists \underline{t} \,:\, C_t < t^{\frac{1}{\gamma-1}} \;\; \forall t > \underline{t}$.*
If in addition η_x is upper-bounded, then $S_\infty = 0$.

Proof. See Appendix A.5. □

Observe that, similar to $R(\cdot)$, $\tilde{R}(k)$ is the long-run ratio of the marginal value of lowering risk to the marginal value of increasing consumption when

$$C_t \propto t^{\frac{k}{\gamma-1}}. \tag{30}$$

If $\tilde{R}(k) < 1$ on this consumption path, for some $k > 1$, then on this path consumption grows too slowly. It is eventually preferable to raise x_t above its implied level of approximately $t^{\frac{k}{\gamma-1}}/A_t$. So if $\lim_{k\downarrow 1} \tilde{R}(k) < 1$, C_t eventually grows more quickly than (30) for some $k > 1$ on the optimal path. Conversely, if $\lim_{k\uparrow 1} \tilde{R}(k) \geq 1$, C_t eventually grows more slowly than (30) for $k = 1$.

If C_t grows more quickly than (30) for some $k > 1$, then the left-hand side of (28) falls more quickly than t^{-k} for some $k > 1$. So $\frac{\partial \delta}{\partial x} x$ does as well. Recalling that $\delta < \frac{\partial \delta}{\partial x} x$, this ensures a positive probability of survival.

If C_t grows more slowly than (30) for $k = 1$, then the left-hand side of (28) falls more slowly than $1/t$. The right-hand side equals $\frac{\partial \delta}{\partial x} x \cdot v = \eta_x / \delta \cdot v$. If η_x is upper-bounded, δ falls more slowly than $1/t$. Cumulative risk is therefore infinite, and survival is impossible.

For illustration, let us evaluate the constant elasticity hazard function of Section 2 for the case of $A_t = e^{gt}$, $g > 0$.

$$\begin{aligned} \tilde{R}(k) &= \lim_{t\to\infty} \bar{\delta} e^{\alpha g t} \beta \left(\frac{t^{\frac{k}{\gamma-1}}}{e^{gt}} \right)^{\beta-1} \frac{t^{\frac{k\gamma}{\gamma-1}}}{e^{gt}} \bar{v} \\ &= \bar{\delta} \beta \bar{v} \lim_{t\to\infty} e^{-(\beta-\alpha)gt} t^{\frac{\beta+\gamma-1}{\gamma-1}k} = 0 \end{aligned} \tag{31}$$

for any k, since $\beta > \alpha$. So $\lim_{k\downarrow 1} \tilde{R}(k) = 0 < 1$. Part (a) of Proposition 5 thus confirms our earlier conclusion that, with hazard function (2), consumption grows at least as quickly as a sufficient power function (in fact it grows exponentially) and that there is a positive probability of survival.

By contrast, consider the constant elasticity hazard function but with $\alpha = \beta$. In this case, (31) $= \infty$ for any k, so $\lim_{k\uparrow 1} \tilde{R}(k) = \infty > 1$. Also, η_x is constant at β, and so upper-bounded. $\delta(A, x) = Ax$ is thus an example of a hazard function satisfying D1–D5 for which the probability of survival on the optimal path is zero (and in fact is so for any $A(\cdot)$ that is eventually bounded above zero).

5.4 Acceleration weakly lowers risk

For any admissible hazard function, the lessons of Section 4 are essentially maintained. The effect of a temporary level effect on the probability of survival is ambiguous. However, if the probability of survival is positive on the planner-optimal policy path, given the baseline technology path, then an acceleration to technological development increases the probability of survival. If the probability of survival is zero on the planner-optimal policy path, then an acceleration to technological development may increase the probability of survival or have no effect.

Acceleration is formally defined and analyzed below. First, we will briefly note the impacts of a marginal, temporary level effect.

Let

$$\eta_{xy}(A_t, x_t) \equiv \frac{\partial}{\partial y}\Big(\frac{\partial \delta}{\partial x}(A_t, x_t)\Big) \cdot \frac{\frac{\partial \delta}{\partial x}(A_t, x_t)}{y_t},$$

for $y \in x, A$, denote the elasticity of $\partial\delta/\partial x$ with respect to y.

If $A_t^{1-\gamma} > \frac{\partial \delta}{\partial x}(A_t, 1)v_t$, so that $x_t = 1$ and the $x_t \leq 1$ constraint binds, then multiplying A_t by m slightly above 1 multiplies δ_t by approximately $m^{\eta_A(A_t,1)} \geq 1$.

If the $x_t \leq 1$ constraint does not bind, so that (28) is maintained with equality as A_t rises, then multiplying A_t by m slightly above 1 has a direct impact and possibly an indirect impact on the hazard rate. The direct impact is again to multiply δ_t by approximately $m^{\eta_A(A_t,x_t)}$. The possible indirect impact is to affect the choice of x_t. Letting $\xi(A_t, x_t)$ denote the elasticity of chosen x to A around (A_t, x_t), to maintain equality (28) as A_t varies we must have

$$\begin{aligned} \xi(A_t, x_t) &= \frac{1-\gamma}{\gamma} - \frac{1}{\gamma}\Big(\eta_{xA}(A_t, x_t) + \xi(A_t, x_t)\eta_{xx}(A_t, x_t)\Big) \\ \implies \xi(A_t, x_t) &= -\frac{\eta_{xA}(A_t, x_t) + \gamma - 1}{\eta_{xx}(A_t, x_t) + \gamma}. \end{aligned}$$

(Observe that v_t is unaffected by a one-period change to A_t and x_t.) If $\xi(A_t, x_t) > 0$ and $x_t = 1$, the marginal increase to A_t does not affect the chosen x_t. Otherwise, the overall elasticity of the hazard rate to A_t, in the context of an instantaneous level effect, is not $\eta_A(A_t, x_t)$ but

$$\eta_A(A_t, x_t) + \xi(A_t, x_t)\eta_x(A_t, x_t). \tag{32}$$

This is negative in the context of hazard function (2), yielding the earlier result that when $x < 1$, temporary level effects lower the hazard rate. Under the weaker conditions here, the sign of (32) is ambiguous. This is illustrated in the beginning of the next subsection, with a hazard function under which, early in time, increases to A—even combined with increases to v—motivate such slow decreases to x than on balance the hazard rate rises.

By contrast, some lasting shocks to the growth path have more predictable effects. Given a growth path $A(\cdot)$ satisfying A1, A2, and A4, we will say that a continuously differentiable (A1-satisfying) growth path $\tilde{A}(\cdot)$ with $\tilde{A}(0) = A(0)$ is an <u>acceleration</u> to $A(\cdot)$ if, for some $\tau > 0$, $\tilde{A}'(t) > A'(t)$ for $t \in (0, \tau)$ and

$$\exists m > 0 \ : \ \tilde{A}(t) = A(t+m) \ \ \forall t \geq \tau.$$

Without loss of generality, we are setting the time denoted "0" to be the beginning of the acceleration.

We will say that the acceleration is <u>permanent</u> if $\tau = \infty$ and <u>temporary</u> otherwise.

Proposition 6. ***Acceleration weakly lowers risk***
Given an admissible technology path $A(\cdot)$ and hazard function $\delta(\cdot)$, and a continuous growth path $\tilde{A}(\cdot)$ that is an acceleration to $A(\cdot)$:

a. *If $X(A(\cdot)) < \infty$, then $X(\tilde{A}(\cdot)) < X(A(\cdot))$.*

b. *If $X(A(\cdot)) = \infty$ and the acceleration is temporary, then $X(\tilde{A}(\cdot)) = \infty$. If $X(A(\cdot)) = \infty$ and the acceleration is permanent, then $X(\tilde{A}(\cdot))$ may be finite or infinite.*

Proof. See Appendix A.6. □

The intuition is the same as illustrated in Section 4. Acceleration in effect horizontally rescales all or part of the hazard curve.

Accelerations vs. level effects

Given a growth path $A(\cdot)$ satisfying A1, A2, and A4, say that a continuously differentiable growth path $\tilde{A}(\cdot)$ with $\tilde{A}(0) = A(0)$ is a <u>level effect</u> to $A(\cdot)$ if, for some $\tau > 0$, $\tilde{A}'(t) > A'(t)$ for $t \in (0, \tau)$ and

$$\exists m > 1 \ : \ \tilde{A}(t) = mA(t) \ \ \forall t \geq \tau.$$

When technology growth is exponential, temporary accelerations are equivalent to level effects. Otherwise, they are sometimes distinct.

Unlike temporary accelerations, level effects do not always decrease cumulative risk outside the exponential growth context. Consider for example hazard function (2) with a technology path $A(t)$ that is roughly stagnant for an arbitrarily long period—say, $[t^* - 100, t^* - 1]$—before the regime-change time t^*, and growth exponential outside this window. A brief level effect around $t^* - 100$ can raise the technology level during the long period of stagnation, which non-negligibly raises cumulative risk, while lowering cumulative risk only negligibly by cutting a vertical slice from the hazard curve following $t^* - 1$.

The direction of technical change

At face value, this is a model in which there is a single dimension to technological development. Inventions simply occur in sequence, each of which increases potential consumption but has an idiosyncratic effect on the hazard rate at any given level of consumption. (Recall that we allow $\delta(A, x)$ to decrease in A.) This one-dimensionality may seem unrealistic. In practice, technological development is surely at least somewhat *directed*, with the tradeoffs between consumption and risk in later periods affected by the extent to which policymakers and market participants in earlier periods have supported research into various types of technology. Consider for example the "richer model" of Jones (2016), in which increases in the value of life relative to consumption motivate increases not only in health spending but also in medical R&D.

As with our assumption that the baseline growth rate of technology at each time is exogenous, however, the assumption that the path of technology is exogenous is also essentially without loss of generality. A path of maximum potential consumption levels $\{A_t\}$ and a hazard function $\delta(A, \cdot)$ simply describe a path of possibilities frontiers over time, without embedding any assumptions about how this path of possibilities frontiers is generated. If we posit a wider space of possible production technologies than the sequence adopted on the baseline path, we must clarify that "accelerations" consist of increases to the rate of motion along the baseline path.

Proposition 6 only applies to accelerations in this sense. Subsidizing the development of risky technologies that would not otherwise have been invented, or choosing a technology path on which they are invented sooner than they would have been but risk-decreasing technologies are not, does not necessarily lower cumulative risk.[13]

5.5 Two hazard functions of interest

We will assume throughout this section that technology growth is exponential at rate $g > 0$.

A lower Inada condition on safety

As we have seen, given a constant elasticity hazard function, δ rises as long as it remains optimal to maximize consumption, and falls immediately once it becomes optimal to begin choosing sub-maximal consumption out of concern for safety. And as noted at the end of Section 3, this result is arguably at odds with the experience of the last century. We will therefore here explore how to tweak the hazard function so that the Kuznets curve is smoothed, and the policy choice variable falls even early in time while the hazard rate is still rising.

A constant elasticity hazard function generates a distinct pair of regimes for the same reason here as in Stokey (1998): namely because, when $x = 1$, marginal "safety expenditures"—decreases to x—produce only finite marginal benefits. That is, there is no "lower Inada condition on safety". It is therefore optimal to maximize consumption until the marginal utility of consumption has fallen and the marginal value of existential risk reduction efforts have risen, as we have seen, and then at once to begin lowering x roughly exponentially. We will say that a hazard function exhibits a lower Inada condition on safety if $\lim_{x\to 1} \frac{\partial \delta}{\partial x} = \infty$. Under this condition, it is opti-

[13]In addition to modeling the policy choice about how much consumption to sacrifice for an instantaneous reduction to the hazard rate, an earlier version of this paper models the technology path as directed by policy as well. The growth model is semi-endogenous, so total potential technology growth is driven by exogenous population growth, but research is optimally allocated between risk-increasing "consumption technology" and risk-decreasing "safety technology". Conceptually, that model sheds light on the same question as this one—how acceleration affects cumulative risk, given an endogenous policy response—but the objects of study are accelerations to population rather than to technology itself. Numerical estimation suggests that acceleration decreases cumulative risk in that context as well, for the same reasons as it does here. When population growth is accelerated, and labor is allocated optimally across fields, society traverses roughly the same technology path but more quickly.

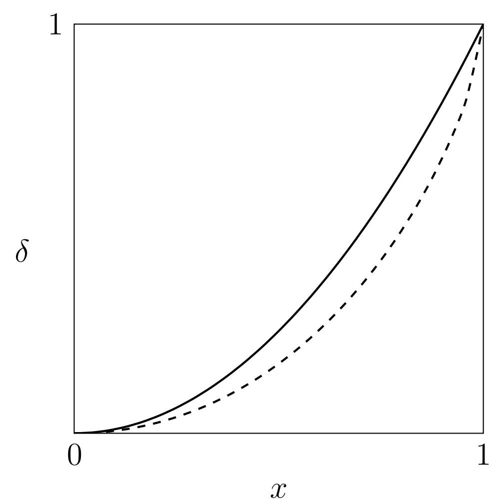
mal to set $x_t < 1$ as long as $v_t > 0$: as long as civilization is worth preserving at all, some expenditures on existential risk reduction are worthwhile.

Not every hazard function with a lower Inada condition on safety behaves like a smoothed version of a constant elasticity hazard function. If the inverse of the hazard function is too concave around $x = 1$ (when A is low), then x may fall rapidly, rather than mildly, from the outset, yielding no early period during which $x \approx 1$. If it is not concave enough around $x = 1$, on the other hand, then early decreases to x produce significant decreases to δ, so that the hazard rate falls even early in time.

One class of hazard functions with the desired features is

$$\delta_t = \bar{\delta} A_t^{\alpha} x_t^{\beta} \frac{1 - (1 - x_t)^{\epsilon}}{x_t}, \quad \epsilon \in \Big(\frac{1}{2}, 1\Big), \tag{33}$$

where the conditions on parameters other than ϵ are as before. The distinction between the hazard functions is illustrated below for the case of $\bar{\delta} A^{\alpha} = 1$, $\epsilon = 0.6$, $\beta = 2$. The solid curve represents the old hazard function; the dashed curve represents the new hazard function, vertical at $x = 1$.



Note that

$$\lim_{x \to 0} \frac{1 - (1 - x_t)^{\epsilon}}{x_t} = \epsilon,$$

so the asymptotics in this case are identical to those in the case of a constant elasticity hazard function (except that the hazard rate is multiplied by ϵ).

The transition dynamics, however, are qualitatively different. Though it is now optimal to set $x < 1$ as long as $v > 0$, x now falls smoothly and δ smoothly rises and falls. The paths of the hazard rate and policy choice are illustrated below for $\epsilon = 0.6$, $A_0 = 2.03$, and otherwise the same parameter values as in Table 1.[14]

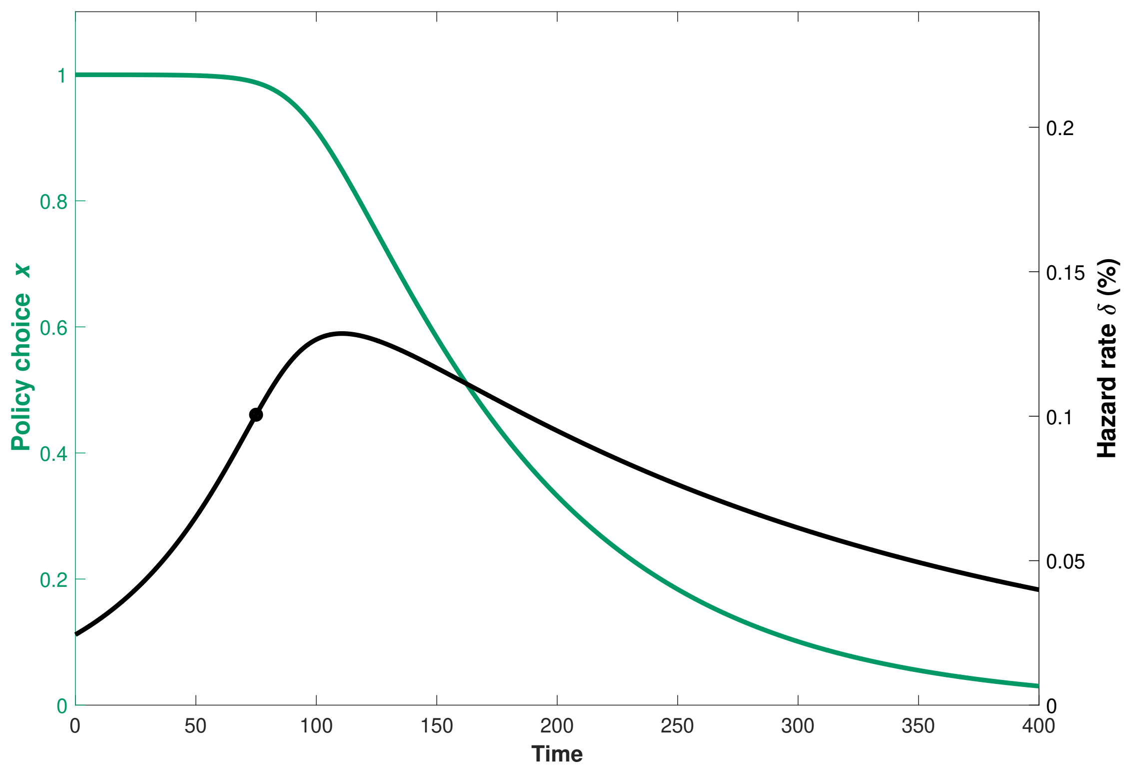


Figure 3: Evolution of the policy choice and the hazard rate along the optimal path given a lower Inada condition on safety expenditure

Derivations and code for replicating the simulation may be found in Appendix B.

[14] A_0 is raised slightly in order to maintain that the value of a statistical life-year "today" (at $t = 75$) is four times per capita consumption, and the hazard rate is approximately 0.1%, despite the fact that, in this model, consumption and the hazard rate are slightly less than maximal even early in time.

Safety in redundancy

The constant elasticity hazard function of Sections 2–4, and its tweak just above, were chosen for clarity. We might however be interested in a better-founded story about the shape of the hazard function, in which the hazard rate is determined by the production of consumption goods and safety goods. For illustration, one relatively straightforward story would be as follows.

- Each unit of consumption (still produced as $C_t = A_t x_t$) poses some risk p of catastrophe per period in the absence of any safety measures.
- For each unit of the consumption good, if one unit of the safety good (produced as $H_t = A_t(1-x_t)$) is allocated to preventing the production process from causing a catastrophe, this fails to prevent a catastrophe with probability $\tilde{b} < 1$. That is, one unit of H per unit of C multiplies the risk posed by each unit of C by $\tilde{b}$, from the baseline of p.
- The probability that the production of a given unit of consumption results in a catastrophe is the probability that (a) there would have been a catastrophe in the absence of any safety measures and (b) all H/C safety measures fail independently: $p\tilde{b}^{H/C}$.
- The probability that the world survives a given period is the probability that all C units of consumption, independently, do *not* generate a catastrophe: $(1 - p\tilde{b}^{H/C})^C$.

In discrete time, the story above would correspond to the hazard function

$$\delta(A_t, x_t) = 1 - \left(1 - p\tilde{b}^{\frac{1-x_t}{x_t}}\right)^{A_t x_t}, \quad \tilde{b} \in (0,1). \tag{34}$$

The continuous-time analog to (34) is

$$\delta(A_t, x_t) = A_t x_t e^{-b\frac{1-x_t}{x_t}}, \quad b > 0 \tag{35}$$

(see Appendix A.7).

Since hazard function (35) lacks any sort of lower Inada condition on $1-x$, x is fixed at 1, and δ rises, early in time while $v > 0$. After the relevant calculations, Propositions 3–5 tell us that (35) yields a Kuznets curve, with δ eventually falling quickly enough to permit survival.

Proposition 7. ***Long-run policy choice and risk given safety in redundancy***
Given hazard function (35), the optimal path features

$$\lim_{t\to\infty} x_t t = \frac{b}{g\gamma}, \tag{36}$$

$$\lim_{t\to\infty} g_{\delta t} = -g(\gamma - 1). \tag{37}$$

Proof. See Appendix A.8. □

Thus the decline in policy choice here is slower than in the constant elasticities case: x declines proportionally to $1/t$ rather than exponentially. This results from the fact that a model of redundancy yields a hazard rate that falls rapidly in the policy choice variable: unit decreases in $A_t x_t$, rather than merely proportional increases, generate proportional decreases to δ. In both cases, however, $x_t \to 0$. And in both cases, δ_t declines exponentially, and so quickly enough to permit survival.

Comparing (37) to the limiting expression for g_δ from Proposition 1, we see that, in the limit, the hazard rate declines more quickly in the redundancy-based model than in the basic model. Mathematically, this follows from the fact that the extra coefficient on $g(\gamma - 1)$ in the limiting expression for g_δ from Proposition 1 is less than one:

$$\alpha > 0, \gamma > 1 \quad \implies \quad \frac{\beta - \alpha}{\beta + \gamma - 1} < 1.$$

Intuitively, this too stems from the fact that, in a redundancy-based model, smaller sacrifices in consumption (linear rather than proportional) are necessary to yield proportional decreases to the hazard rate. The planner's response to this expanded possibilities frontier comes partially in the form of slower increases in foregone consumption, as described by (36), and partially in the form of faster declines in the hazard rate, as described by (37).

6 Transition risk

A hazard function of the form $\delta(A_t, x_t)$ captures what might be called "state risk": δ depends on the *level* of technology. On this framing, it is perhaps unsurprising that escaping risky states more quickly lowers cumulative risk.

But risk may instead be "transitional": posed by *technological development*. This is the intuition captured by Jones's (2016) "Russian roulette" model of technological development and (2023) model of AI risk, and by Bostrom's (2019) analogy to drawing potentially destructive balls from an urn. Perhaps stagnation at a given level of technology is essentially safe, and risk arises in the process of discovering and deploying new technologies with unknown consequences.

6.1 A transition-risk-based hazard function

To explore this possibility, suppose δ increases in $\dot{A}_t$ instead of, or as well as, in A_t. For simplicity, we will again restrict our consideration to a constant elasticity hazard function:

$$\delta_t = \bar{\delta} A_t^{\alpha} \dot{A}_t^{\zeta} x_t^{\beta}, \quad \bar{\delta} > 0,\ \zeta \geq 0,\ \beta > \alpha + \zeta > 0,\ \beta > 1. \tag{38}$$

We will also again assume that A grows at a constant exponential rate $g > 0$ on the baseline technology path.

Since $g_{\dot{A}} = g$ on this path, $\beta > \alpha + \zeta$ is now the condition necessary for survival without $C_t = A_t x_t \to 0$, and $\alpha + \zeta > 0$ is now the condition under which growth increases the hazard rate when x is fixed.

Our original hazard function (2) is the special case of (38) with $\zeta = 0$. This model is thus an alternative generalization of hazard function (2), complementary to that of Section 5.

As long as $\zeta > 0$, however, the model is most straightforwardly interpreted as one in which new technologies—new "draws from Bostrom's urn"—consist of absolute increases to A. The introduction of multiple technologies can pose more, less, or equal risk if they are introduced concurrently than if they are introduced in sequence, depending on whether ζ is greater than, less than, or equal to 1. The development of more advanced technologies can pose more, less, or equal risk as compared to the development of less advanced technologies, depending on the sign of α.

Alternatively, the model may be interpreted as one in which new technologies consist of proportional increases to A. This can be seen by rewriting the hazard function as

$$\delta_t = \bar{\delta} A_t^{\alpha+\zeta} \left(\frac{\dot{A}_t}{A_t}\right)^{\zeta} x_t^{\beta}.$$

On this interpretation, the assumption that $\alpha + \zeta > 0$ amounts to the assumption that the development of more advanced technologies poses more risk than the development of less advanced technologies. Because $\dot{A}/A$ has been approximately constant throughout the last century, the view that the hazard rate has risen must be attributed to the increasing danger of each "technological development" in this sense.

6.2 Acceleration still typically weakly lowers risk

Since $\dot{A}$ is proportional to A, the planner's problem is unchanged (up to a coefficient g^{ζ} that can be rolled into $\bar{\delta}$). Baseline x and δ paths, and S_∞, are unchanged. Let A^* denote the uppermost technology level at which it is optimal to set $x = 1$.

To examine the impact of accelerating growth on cumulative risk, it will again be helpful to integrate the hazard curve with respect to A:

$$\begin{aligned} \int_0^\infty \bar{\delta} A_t^\alpha \dot{A}_t^\zeta x_t^\beta dt &= \int_{A_0}^\infty \bar{\delta}\, A^\alpha\, \dot{A}_A^\zeta\, x_A^\beta\, dA \left(\frac{dA}{dt}\right)^{-1} \\ &= \int_{A_0}^\infty \bar{\delta}\, A^\alpha \dot{A}_A^{\zeta-1}\, x_A^\beta\, dA. \end{aligned} \tag{39}$$

Writing x as a function of A, we have

$$x_A = \begin{cases} 1 & A \leq A^*, \\ \left(\bar{\delta}\beta A^{\alpha+\gamma-1}\dot{A}_A^\zeta v_A\right)^{-\frac{1}{\beta+\gamma-1}} & A > A^*. \end{cases} \tag{40}$$

Substituting (40) into (39) yields

$$\begin{aligned} &\int_{A_0}^{A^*} \bar{\delta} A^\alpha \dot{A}_A^{\zeta-1} dA \\ &+ \int_{A^*}^\infty \left(\bar{\delta}^{1-\gamma}\beta^\beta A^{(\beta-\alpha)(\gamma-1)} v_A^\beta\right)^{-\frac{1}{\beta+\gamma-1}} \dot{A}_A^{\zeta\frac{\gamma-1}{\beta+\gamma-1}-1} dA. \end{aligned} \tag{41}$$

Recall that temporary accelerations, here equivalent to permanent level effects, are temporary increases to $\dot{A}$. While they are underway, they increase v_A, and they have no impact on v_A for technology levels after they conclude.

Before A^*, therefore, temporary accelerations decrease cumulative risk if $\zeta < 1$, increase it if $\zeta > 1$, and have no effect if $\zeta = 1$. The $\zeta = 1$ case is arguably central, especially before A^*: if no effort is made to mitigate the dangers of risky experiments, it should not matter whether they occur serially or in parallel.

After A^*, however, temporary accelerations decrease cumulative risk as long as

$$\zeta \frac{\gamma - 1}{\beta + \gamma - 1} \leq 1. \tag{42}$$

It is sufficient, though not necessary, for (42) that

$$\zeta \leq 1 \quad \text{or} \quad \alpha \geq -1,\ \gamma \leq 2.$$

The $\zeta \leq 1$ case follows from the fact that $\frac{\gamma-1}{\beta+\gamma-1} < 1$. The $\alpha \geq -1$, $\gamma \leq 2$ case follows from the fact that if $\alpha \geq -1$, then $\zeta < \beta + 1$, so $\frac{\zeta}{\beta+1} < 1$. Since macroeconomic estimates of $\gamma \leq 2$ are standard, this result suggests that, for typical parameter values, accelerations to technology growth lower or have no effect on cumulative risk on the optimal path in the context of transition risk.

The condition of an upper bound on γ may be counterintuitive, because higher values of γ lower the marginal utility of consumption and motivate more rapid reallocations of resources from consumption to safety. The result is driven by the fact that, when γ is high, the marginal utility of consumption rises rapidly as x is cut, so only a small cut to x suffices to maintain the condition that the marginal utility of consumption equals the marginal disutility incurred by raising the hazard rate. The higher γ is, the more quickly x falls as A rises, but the *less* sensitive x is to a change in $\partial\delta/\partial x$ at a given value of A.

Level effects that cross the A^* threshold will be ignored for simplicity.

Growth effects are simply permanent increases to $\dot{A}$. If they begin at $\underline{A}$, they increase v_A for all $A \geq \underline{A}$. This shrinks the hazard rate at all $A > \max(\underline{A}, A^*)$. If $\underline{A} < A^*$, this also lowers A^*. Both effects shrink cumulative risk (41) and so raise S_∞. Otherwise their impacts are like those of level effects.

Stagnation vs. deceleration

When $\zeta > 0$, complete stagnation ($\dot{A} = 0$) is the safest path of all. Nevertheless, we have seen that given a positive growth rate, faster growth often decreases risk.

The key to this puzzle is that, given stagnation at $\bar{A}$, levels of $A > \bar{A}$ are never attained. Cumulative risk is therefore not (39) but (39) with the ∞ replaced with $\bar{A}$. Absent stagnation, however slow the growth rate, all levels of A are attained. The growth rate only determines the risk endured at each one. The direct cost of faster progress during a given range of A-values (higher risk per unit time, to the extent that $\zeta > 0$) is partially, and may be more than fully, outweighed by the fact that faster progress motivates more mitigation at each point in time, in combination with the now familiar fact that when progress is faster we do not linger in a given range of A-values as long.

7 Conclusion

Human activity can create or mitigate existential risks. The framework presented here illustrates that, under relatively mild assumptions, existential risk satisfies the conditions that give rise to a Kuznets curve. This observation offers a potential economic explanation for the claim by some prominent thinkers that humanity is in a critical "time of perils". We may be economically advanced enough to be able to destroy ourselves, but not yet enough that we are willing to make large sacrifices for the sake of safety. If we are indeed living through the time of perils, reductions to existential risk today will have a massive expected impact on the course of the long-run future.

At the same time, this framework highlights a channel through which some efforts intended to reduce existential risk may backfire. When technology is efficiently regulated, even by a policymaker with little concern for the long-term future, broad-based decelerations to technological development generally either worsen or have no impact on the odds of long-term survival. This cost can be significant, with proportional consumption decreases having comparable impacts to proportional increases in the planner's rate of time preference. In the extreme, permanent technological stagnation can make a catastrophe inevitable that might otherwise have been avoided.

This is far from an argument against regulating the use of risky technologies. Indeed, the primary channel explored here through which technological development lowers risk is that it hastens the day when regulation is severe. Some recent reactions against calls to heavily regulate AI, e.g. that of Andreessen (2023), might be read as expressing the view that our "x" should never be set far below one. If that is so, it is not for the reasons presented in this paper.

The reasoning presented here also does not imply that decelerations to technological development inevitably raise cumulative risk: only that they typically do so when technology is efficiently regulated. If the policy response to dangerous new technologies is not efficient—for instance, if there is a long-term limit to the speed at which new safeguards can be imposed—then the impact of a technological acceleration on cumulative risk is ambiguous. In fact, Shulman and Thornley (2024) argue that the policy response to hazardous technologies to date has been far from efficient. The appropriate lesson is only that, to the extent that the regulatory regime does *or will eventually* move toward equating the marginal utility of consumption to the marginal discounted utility of existential risk reduction (per unit of consumption sacrificed), consumption-increasing technological development today has the unseen but potentially large benefit of speeding future safety efforts. For slowing technological development to lower cumulative risk, the policy inefficiency in question must be severe and lasting enough to outweigh this benefit.

In this light, further research on the nature of any policy distortions around the regulation of hazardous technologies would be valuable. Exploring the long-term implications of other models of optimal policy in the face of anthropogenic existential risk—beyond the simple state- and transition-risk-based relationships explored here—could be valuable as well, so as to better characterize the scope of the result that efficiently regulated acceleration weakly lowers cumulative risk. If plausible models are found under which the result is overturned, this will naturally pose an important question which can only be answered empirically. For now, however, the results presented here suggest that even those exclusively concerned with reducing cumulative existential risk should often cheer technological advances despite their short-term hazards, and advocate risk-reduction measures today only when they are sufficiently targeted and the costs to technological development are sufficiently small.

References

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

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.

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

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

Brock, William A. and M. Scott Taylor, "Economic Growth and the Environment: A Review of Theory and Empirics," in Philippe Aghion and Steven N. Durlauf, eds., *The Handbook of Economic Growth*, Vol. 1, Elsevier, 2005, chapter 28, pp. 1749–1821.

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

Chetty, Raj, "A Bound on Risk Aversion Using Labor Supply Elasticities," *American Economic Review*, 2006, *96* (5).

Cotton-Barratt, Owen, "Allocating risk mitigation across time," 2015. Future of Humanity Institute Technical Report #2015-2.

Cowen, Tyler and Derek Parfit, "Against the Social Discount Rate," in Peter Laslett and James S. Fishkin, eds., *Justice Between Age Groups and Generations*, New Haven: Yale University Press, 1992, pp. 144–161.

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. `https://futureoflife.org/open-letter/pause-giant-ai-experiments/`.

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

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

—, "The A.I. Dilemma: Growth Versus Existential Risk," 2023.

Klenow, Peter J., Charles I. Jones, Mark Bils, and Mohamad Adhami, "Population and Welfare: The Greatest Good for the Greatest Number," December 2023.

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.

MacAskill, William, *What We Owe the Future: A Million-Year View*, Oneworld Publications, 2022.

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.

Meadows, Donella H., Dennis L. Meadows, Jørgen Randers, and William W. Behrens III, *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*, New York: Universe Books, 1972.

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

Moynihan, Thomas, *X-Risk: How Humanity Discovered Its Own Extinction*, Urbanomic, 2020.

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

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

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

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

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.

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.

Appendices

A Proofs

A.1 Characterizing the optimal path

Necessary and sufficient conditions

The dynamic optimization problems analyzed in this paper all feature one choice variable x and one state variable S. Expected flow utility at t is $S_t u(A_t, x_t)$ for a continuously differentiable function $u(\cdot)$, strictly concave in x, with a lower Inada condition on x. The law of motion for S is given by $-S_t\delta(A_t, \dot{A}_t, x_t)$ for a continuously differentiable function $\delta(\cdot)$. A and $\dot{A}$ are independent of x, so operate simply as functions of t. Letting v denote the costate variable on S, the current value Lagrangian corresponding to the problem is then

$$\mathcal{L}(S_t, x_t, v_t, \mu_t, t) = S_t u(x_t, t) - v_t S_t \delta(x_t, t) + \mu_t S_t (1 - x_t) \tag{43}$$

(abusing notation slightly by reusing $u(\cdot)$ and $\delta(\cdot)$ as functions of time), where μ_t represents the the Lagrange multiplier on x_t. We impose the $x_t \leq 1$ constraint but not the $x_t \geq 0$ constraint because the latter can never bind, by the lower Inada condition on $u(\cdot)$.

(43) satisfies the Mangasarian concavity condition that $\mathcal{L}(\cdot)$ is everywhere concave in S and x. So, applying Caputo (2005), Theorems 14.3-4 and Lemma 14.1,[15] given continuous paths of $x \in [0, 1]$ and $S \in [0, 1]$ with $S_0 = 1$ and $\dot{S}_t = -S_t\delta(x_t, t)$, we have that the x, S path is optimal if—and, given piecewise continuity of x and S, only if—for some piecewise differentiable path of v and some piecewise continuous path of $\mu \geq 0$, at all t the following first-order conditions are satisfied

$$\frac{\partial \mathcal{L}}{\partial x_t}(S_t, x_t, v_t, \mu_t, t) = 0, \tag{44}$$

$$\frac{\partial \mathcal{L}}{\partial \mu_t}(S_t, x_t, v_t, \mu_t, t) \geq 0, \tag{45}$$

$$\mu_t \frac{\partial \mathcal{L}}{\partial \mu_t}(S_t, x_t, v_t, \mu_t, t) = 0 \tag{46}$$

[15] 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.

as well as the transversality condition that

$$\lim_{t\to\infty} e^{-\rho t} v_t = \lim_{t\to\infty} e^{-\rho t} v_t S_t = 0. \tag{47}$$

Furthermore, given optimal paths of x and S and corresponding paths of v and μ, v will satisfy

$$\begin{aligned} \dot{v}_t &= \rho v_t - u(x_t, t) - v_t \dot{S}_t \\ &= (\rho + \delta(x_t, t)) v_t - u(x_t, t) \end{aligned} \tag{48}$$

except at any discontinuity points of x, at which v will have different right and left derivatives.

Interpreting the transversality condition

Given a continuous v path, only the paths of x and μ defined by

$$x_t = \begin{cases} 1, & \frac{\partial u}{\partial x}(1,t) - \frac{\partial \delta}{\partial x}(1,t) v_t \geq 0; \\ x_t \,:\, \frac{\partial u}{\partial x}(x_t,t) - \frac{\partial \delta}{\partial x}(x_t,t) v_t = 0, & \text{otherwise} \end{cases} \tag{49}$$

$$\mu_t = \frac{\partial u}{\partial x_t}(x_t, t) - \frac{\partial \delta}{\partial x_t}(x_t, t) v_t \tag{50}$$

satisfy (44)–(46) for all t. Any such x path is well-defined, by the continuous differentiability of $u(\cdot)$ and $\delta(\cdot)$ and the fact that $u(\cdot)$ and $\delta(\cdot)$ strictly increase in x. Any such x path is also continuous in time, by the twice continuous differentiability of $u(\cdot)$ and $\delta(\cdot)$ (expressed as functions of x and A) and the continuous differentiability of $A(\cdot)$, and the implicit function theorem. Any such μ path is then also continuous in time by the composition of continuous functions. To show there exists an optimal path, and that only one such path is piecewise continuous, it will now suffice to show that there is a unique v path for which (47)–(48) are satisfied given the corresponding x path (49) and its implied S path, and that the corresponding x path is piecewise continuous (in fact it is continuous everywhere).

The solution to differential equation (48) is

$$v_t = e^{\int_0^t (\rho + \delta_s) ds} \Big(v_0 - \int_0^t e^{-\int_0^s (\rho + \delta_q) dq} u(x_s, s) ds \Big) \tag{51}$$

$$\implies v_0 = \int_0^t e^{-\rho s} S_s u(x_s, s) ds + e^{-\rho t} S_t v_t. \tag{52}$$

Since (52) is continuous in t (by the continuity of x in t and the continuous evolution of S) and holds for all t, v satisfies (47)–(48) iff

$$v_0 = \int_0^\infty e^{-\rho t} S_t u(x_t, t) dt. \tag{53}$$

That is, the value of increasing the probability of survival, as of time 0, must equal the expected utility of the future (given survival past time 0).

Given (49), v_t determines x_t for all t, and given (48), v_t and x_t determine $\dot{v}_t$ for all t. For a given v_0, therefore, there is a unique path of v—and thus of x, and thus of S—compatible with (48)–(49). We will now show that there is at least one value of v_0 for which (53) is satisfied, given the corresponding x and S paths. For such a v_0, the corresponding variable paths will by construction satisfy (44)–(47), and thus constitute an optimum.

Existence

Let $v(v_0)$ and $x(v_0)$ denote the unique paths of v and x compatible with (48)–(49) for which $v_0(v_0) = v_0$. By (51), $\lim_{v_0 \to -\infty} v_t(v_0) = -\infty$ for all $t \geq 0$. By (49), therefore, for every $t \geq 0$, there is a $\tilde{v}_0$ such that $x_t(v_0) = 1$ for all $v_0 < \tilde{v}_0$. Let $s \geq 0$ denote a time at which $A_s \geq 1$, and choose $\tilde{v}_0$ low enough that $\tilde{v}_s < 0$ and thus $x_s(\tilde{v}_0) = 1$. By (48), because $u(1, s) \geq 0$, $\dot{\tilde{v}}_t < 0$. We thus have $\tilde{v}_t < 0$, and thus $x_t = 1$, for all $t \geq s$.

Now 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 with respect to time, there would be a t with $v_t(v_0) = v_t(\tilde{v}_0)$, and (48)–(49) would allow us to unroll the paths identically so as to yield $v_0 = \tilde{v}_0$. Thus, if $v_0 < \tilde{v}_0$, $x_t(v_0) \geq x_t(\tilde{v}_0)$ for all $t \geq 0$. It follows that, for some sufficiently low $\underline{v}_0$, the right-hand side of (53) exceeds the left-hand side.

For every optimization problem under consideration, there is some $\overline{U}$ by which feasible values of the right-hand side of (53) are upper-bounded. So, for $\overline{v}_0 > \overline{U}$, the left-hand side of (53) exceeds the right-hand side.

The implicit function theorem gives us that x_t is continuous in v_t. (48) then implies that $\dot{v}_t$ is continuous in v_t for all t, and thus that $v_t(v_0)$, then $x_t(v_0)$, and then ultimately the right-hand side of (53) are is continuous in v_0 for all t. It follows from the intermediate value theorem that there exists a $v_0 \in (\underline{v}_0, \overline{v}_0)$ for which (53) holds.

Uniqueness

Standard uniqueness results (e.g. Caputo (2005), Theorem 14.4 cited above) do not immediately apply here, because the Lagrangian is linear, not strictly concave, in the state variable S. Fortunately, this can easily be remedied by defining the state variable to be e.g. S^2 without affecting any conditions necessary for the other results.

Uniqueness (among piecewise continuous x paths) also follows immediately from the observations that a path is optimal iff v_0 attains its maximum feasible value and that, given (44)–(47), v_0 determines a unique path for every variable.

A.2 Long-run g_v and proof of Proposition 2

Long-run constancy of g_v for all γ

Observe from (48) that, because v is the costate variable on S, it must follow the law of motion

$$\begin{aligned} \dot{v}_t &= (\rho + \delta_t)v_t - u(C_t) \\ \implies g_{vt} &= \rho + \delta(A_t, x_t) - \frac{u(A_t x_t)}{v_t}. \end{aligned} \tag{54}$$

Let

$$\tilde{\beta} \equiv \beta + \gamma - 1.$$

From (13), once x_t is interior we have

$$x_t = A_t^{-\frac{\alpha+\gamma-1}{\tilde{\beta}}} \left(\bar{\delta}\beta v_t\right)^{-\frac{1}{\tilde{\beta}}}. \tag{55}$$

Substituting (55) into (54) yields

$$g_{vt} = g_v(v_t, t) \equiv \begin{cases} \rho + K A_t^{\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}} v_t^{-\frac{\beta}{\tilde{\beta}}} + \frac{1}{1-\gamma} v_t^{-1}, & \gamma \neq 1; \\ \rho + \log\left(A_t^{-\frac{\beta-\alpha}{\beta}} \left(\bar{\delta}\beta v_t\right)^{-\frac{1}{\beta}}\right) v_t^{-1}, & \gamma = 1, \end{cases} \tag{56}$$

where

$$K \equiv \bar{\delta}^{-\frac{1-\gamma}{\tilde{\beta}}} \left(\beta^{-\frac{\beta}{\tilde{\beta}}} - \frac{1}{1-\gamma} \beta^{-\frac{1-\gamma}{\tilde{\beta}}}\right).$$

If $\gamma > 1$, recalling that v_t monotonically increases and that $A_t \to \infty$, the central term of (56) vanishes. Also, in this case, v is upper-bounded, so it approaches an upper bound v^* by the monotone convergence theorem. So $\lim_{t\to\infty} g_{vt}$ is defined, with

$$\lim_{t\to\infty} g_{vt} = \rho + \frac{1}{v^*(1-\gamma)}. \tag{57}$$

This limit cannot be positive, because v is upper-bounded, and it cannot be negative, because v increases with time. So $\lim_{t\to\infty} g_{vt} = 0$, and $v^* = \frac{1}{\rho(\gamma-1)}$.

If $\gamma < 1$, then $K < 0$, and the central term of (56) grows in magnitude without bound, fixing v. v must therefore also grow without bound, or else g_{vt} is eventually negative.

Now observe that

$$\begin{aligned}
\dot{g_{vt}} &= KA_t^{\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}} v_t^{-\frac{\beta}{\tilde{\beta}}}\Big(\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}g - \frac{\beta}{\tilde{\beta}}g_{vt}\Big) - \frac{1/v_t}{1-\gamma}g_{vt} \\
&= \Big(g_{vt} - \rho - \frac{1/v_t}{1-\gamma}\Big)\Big(\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}g - \frac{\beta}{\tilde{\beta}}g_{vt}\Big) - \frac{1/v_t}{1-\gamma}g_{vt} \\
&= -\frac{\beta}{\tilde{\beta}}g_{vt}^2 + \Big(\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}g + \frac{\beta}{\tilde{\beta}}\rho + \frac{1}{\tilde{\beta}v_t}\Big)g_{vt} - \Big(\rho + \frac{1/v_t}{1-\gamma}\Big)\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}g.
\end{aligned}$$

This differential equation has two steady states, both positive. Since $1/v_t \to 0$, the quadratic formula tells us that these steady states approach ρ and $g(\beta-\alpha)(1-\gamma)/\beta$, with the former attractive and the latter repulsive. By (19), ρ is higher, and is ruled out as a steady state by the transversality condition (47). Then because the limits

$$\begin{aligned}
&\lim_{t\to\infty} \dot{g_v}(g_v, t) > 0 \;\; \forall g_v \in \Big(\frac{(\beta-\alpha)(1-\gamma)}{\beta}g, \rho\Big), \\
&\lim_{t\to\infty} \dot{g_v}(g_v, t) < 0 \;\; \forall g_v < \frac{(\beta-\alpha)(1-\gamma)}{\beta}
\end{aligned}$$

are defined and continuous in g_v, we must have

$$\lim_{t\to\infty} g_{vt} = \frac{(\beta-\alpha)(1-\gamma)}{\beta}g. \tag{58}$$

Otherwise we would have $g_v \to -\infty$, ruled out by the monotonicity of v, or $g_v \to \rho$, ruled out above.

The $\gamma = 1$ case is analogous to the $\gamma > 1$ case. Differentiating (56) with respect to time yields $\dot{g_{vt}}$ strictly and continuously increasing in g_{vt} from $-\infty$ at $v_t = 0$ to ρ at $v_t = \infty$. There is thus a unique, positive, and repulsive "time-dependent steady state" value of g_v (i.e. g_v for which $\dot{g_v}(g_v, t) = 0$) which declines to zero as $t \to \infty$. The limits

$$\lim_{t\to\infty} \dot{g_v}(g_v, t) > 0 \ \ \forall g_v > 0,$$
$$\lim_{t\to\infty} \dot{g_v}(g_v, t) < 0 \ \ \forall g_v < 0$$

are defined and continuous in g_v, and we must have

$$\lim_{t\to\infty} g_{vt} = 0$$

to avoid $g_v \to -\infty$ or $g_v \to \infty$.

Proof of Proposition 2

With the limiting behavior of g_v pinned down, the asymptotic behavior of the other variables follows straightforwardly. Substituting (58) for g_{vt} into expression (14) for g_{xt} (and observing that the expression captures all $\gamma \leq 1$) produces

$$\lim_{t\to\infty} g_{xt} = -\frac{\alpha}{\beta} g,$$

and adding αg then produces the limit of $g_{Ax} = g_C$:

$$\lim_{t\to\infty} g_{Ct} = \frac{\beta - \alpha}{\beta} g. \tag{59}$$

For the hazard rate, rearrange (56) to get

$$v_t = \frac{u(C_t)}{\rho + \delta_t - g_{vt}}, \tag{60}$$

and substitute (60) into (23) to get

$$\delta_t = \begin{cases} \frac{\rho+\delta_t-g_{vt}}{\beta} \frac{1-\gamma}{1-C_t^{\gamma-1}}, & \gamma < 1; \\ \frac{\rho+\delta_t-g_{vt}}{\beta \log(C_t)}, & \gamma = 1. \end{cases}$$

Solving for δ_t,

$$\delta_t = \begin{cases} \frac{(\rho - g_{vt})(1-\gamma)}{\beta\left(1-C_t^{\gamma-1}\right)-1+\gamma}, & \gamma < 1; \\ \frac{\rho - g_{vt}}{\beta \log(C_t) - 1}, & \gamma = 1. \end{cases}$$

In the $\gamma < 1$ case, the limit of g_v (58) and $C \to \infty$ from (59) imply

$$\lim_{t\to\infty} \delta_t = \frac{(\rho - (\beta-\alpha)(1-\gamma)g/\beta)(1-\gamma)}{\beta+\gamma-1}.$$

In the $\gamma = 1$ case, substitute 0 for g_{vt} and observe that, by (59),

$$\lim_{t\to\infty} \frac{C_t}{e^{\frac{\beta-\alpha}{\beta}gt}} = \underline{C}$$

for some $\underline{C} > 0$, so that

$$\begin{aligned} \lim_{t\to\infty} \delta_t t &= \lim_{t\to\infty} \frac{\rho - g_{vt}}{\beta\left(\log(C_t/e^{\frac{\beta-\alpha}{\beta}gt}) + \log(e^{\frac{\beta-\alpha}{\beta}gt})\right)/t - 1/t} \\ &= \lim_{t\to\infty} \frac{\rho}{\beta \log(\underline{C})/t + (\beta-\alpha)g - 1/t} \\ &= \frac{\rho}{(\beta-\alpha)g}. \end{aligned}$$

A.3 Proof of Proposition 3

Suppose that $R^* \leq 1$, and, by contradiction, that we do not have $C^* = \infty$.

By the failure of $C^* = \infty$, there is an increasing and unbounded sequence of times, $t_n \to \infty$, such that $C_{t_n} \leq \overline{C}$ $\forall n \geq 1$.

Consider the sequence of consumption levels $n\overline{C}$ $\forall n \geq 1$. Since $n\overline{C} \to \infty$, by $R^* \leq 1$ we have

$$\lim_{n\to\infty} R(n\overline{C}) = \lim_{n\to\infty} \lim_{A\to\infty} \frac{\partial\delta}{\partial x}\left(A, \frac{n\overline{C}}{A}\right) \frac{(n\overline{C})^\gamma}{A\rho(\gamma-1)} \leq 1. \tag{61}$$

By D5, $\frac{\partial\delta}{\partial x}(A,x)$ weakly increases in x for any A. So

$$R(C_{t_n}) \leq R(n\overline{C})\left(\frac{C_{t_n}}{n\overline{C}}\right)^\gamma \leq R(n\overline{C})\, n^{-\gamma} \quad \forall n, \tag{62}$$

where the first inequality follows from the fact that $n\overline{C} \geq C_{t_n}$ for each n, and the second follows from $\overline{C} \geq C_{t_n}$ for each n. By (61), $R(n\overline{C})n^{-\gamma} < 1$ for sufficiently large n, so by (62) and A3, there exists an $\underline{n}$ such that

$$\frac{\partial \delta}{\partial x}\big(A_{t_n}, \frac{C_{t_n}}{A_{t_n}}\big)\frac{C_{t_n}^{\gamma}}{A_{t_n}\rho(\gamma-1)} < 1 \;\; \forall n > \underline{n}.$$

Since v_t cannot exceed $\frac{1}{\rho(\gamma-1)}$,

$$\frac{\partial \delta}{\partial x}\big(A_{t_n}, \frac{C_{t_n}}{A_{t_n}}\big)v_{t_n} < A_{t_n}C_{t_n}^{-\gamma} \;\; \forall n > \underline{n}.$$

This is compatible with optimality only if $x_{t_n} = 1$. But this is impossible for sufficiently large n, since $C_{t_n} = A_{t_n}x_{t_n} \leq \overline{C}$ and $\lim_{n\to\infty} A_{t_n} = \infty$.

Suppose that $R^* > 1$ and, by contradiction, that $C^* = \infty$. Then there is some $\underline{C}$ such that $R(\underline{C}) > 1$:

$$\lim_{A\to\infty} \frac{\partial \delta}{\partial x}\big(A, \frac{\underline{C}}{A}\big)\frac{\underline{C}^{\gamma}}{A\rho(\gamma-1)} > 1.$$

So there is an $\underline{A}$ such that

$$\frac{\partial \delta}{\partial x}\big(A, \frac{\underline{C}}{A}\big)\frac{1}{\rho(\gamma-1)} > A\underline{C}^{-\gamma} \tag{63}$$

for all $A \geq \underline{A}$. Furthermore, because the left-hand side weakly increases in $\underline{C}$ by D5 and the right-hand side strictly decreases in $\underline{C}$, (63) holds for all $A \geq \underline{A}$ and $C \geq \underline{C}$. By A4, and the supposition that $C^* = \infty$, there is a $\underline{t}$ such that

$$\frac{\partial \delta}{\partial x}\big(A_t, \frac{C_t}{A_t}\big)\frac{1}{\rho(\gamma-1)} > A_t C_t^{-\gamma} \;\; \forall t \geq \underline{t}. \tag{64}$$

Finally, optimality requires

$$\begin{aligned} A_t^{1-\gamma}x_t^{-\gamma} &\geq \frac{\partial \delta}{\partial x_t}\big(A_t, x_t\big)v_t \;\; \forall t \\ \implies (A_t x_t)^{1-\gamma}/v_t &\geq \frac{\partial \delta}{\partial x_t}\big(A_t, x_t\big)x_t \geq \delta(A_t, x_t), \end{aligned}$$

with the final inequality holding because, by D5, $\frac{\partial \delta}{\partial x}x \geq \delta$. Given $C^* = \infty$, since v_t is upper-bounded, it follows that $\delta_t \to 0$. With $\delta_t \to 0$ and $C_t \to \infty$, v_t approaches its upper bound of $\frac{1}{\rho(\gamma-1)}$.

It therefore follows from (64) that, for sufficiently large t,

$$\frac{\partial \delta}{\partial x}\big(A_t, \frac{C_t}{A_t}\big) v_t > A_t C_t^{-\gamma}.$$

This is incompatible with optimality. Thus, if $R^* > 1$, it is impossible that $C^* = \infty$.

A.4 Proof of Proposition 4a

It is optimal to set $x_t = 1$ as long as, at $x = 1$, the marginal flow disutility of decreasing x weakly exceeds the marginal expected utility of doing so via decreasing the hazard rate:

$$A_t^{1-\gamma} \geq \frac{\partial \delta}{\partial x}(A_t, 1)\, v_t. \tag{65}$$

It is optimal to set $x_t < 1$ as long as (65) fails, maintaining

$$A_t^{1-\gamma} x_t^{-\gamma} = \frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t \tag{66}$$

$$\implies x_t = A_t^{\frac{1-\gamma}{\gamma}} \Big(\frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t\Big)^{-\frac{1}{\gamma}}. \tag{67}$$

The uniqueness of the optimal path is shown in Appendix A.1.

Proof that $\lim_{t \to -\infty} x_t = 1$

We will show that there exists a time $\underline{t}$ such that $v_{\underline{t}} \leq 0$. It then follows immediately that $x_t = 1$ for $t \leq \underline{t}$.

Let

$$T \equiv A^{-1}\big((\gamma - 1)^{\frac{1}{1-\gamma}}\big)$$

denote the time at which $A_T = (\gamma - 1)^{\frac{1}{1-\gamma}}$, and at which therefore $u(A_T) = -1$. If $v_T \leq 0$, the result follows immediately. Let us therefore assume that $v_T > 0$.

For $t < T$,

$$v_t = \int_t^\infty e^{-\rho(s-t)-\int_t^s \delta_q dq} u(C_s) ds$$
$$= \int_t^T e^{-\rho(s-t)-\int_t^s \delta_q dq} u(C_s) ds \ + \ e^{-\rho(T-t)-\int_t^T \delta_q dq} v_T. \tag{68}$$

Since $u(C_s) \leq u(A_s) \leq -1$ for $s \leq T$, the first term of (68) is negative—indeed, an integral over s of values which are negative for all s. The integral is shrunk in magnitude when, for all s, $u(C_s)$ is replaced with -1 and the discount factor $e^{-\rho(s-t)-\int_t^s \delta_q dq}$ replaced with its minimum value across the range, namely the discount factor at T. So

$$v_t < (t - T + v_T) e^{-\rho(T-t)-\int_t^T \delta_q dq}$$
$$\implies v_{T-v_T} < 0.$$

This proof admittedly "takes the model too literally", in assuming that technology growth has always been exponential and that therefore life was not worth living before some point in the past. Still, the dynamic it bluntly illustrates should not be controversial. When $\gamma > 1$, proportional sacrifices in consumption—decreases to x—carry greater utility costs the lower the baseline consumption level is. Early in time, the discounted value of civilization v and the baseline consumption level A were both low, so large sacrifices for safety would not have been optimal.

Proof that $\lim_{t\to\infty} x_t = 0$ if η_A is bounded above $1 - \gamma$

Generalizing (67), whether or not the $x_t \leq 1$ constraint binds we have

$$x_t \leq A_t^{\frac{1-\gamma}{\gamma}} \left(\frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t \right)^{-\frac{1}{\gamma}}. \tag{69}$$

We will show that if $\eta_A(\cdot)$ is bounded above $1 - \gamma$, the right-hand side has an upper bound which falls to 0 as (by A3) $A_t \to \infty$.

Because by D1 δ is positive, by D2 and D5 we have $\frac{\partial \delta}{\partial x}(A_t, x_t) \geq \delta(A_t, x_t)$. The right-hand side is thus bounded above by

$$A_t^{\frac{1-\gamma}{\gamma}} \left(\delta(A_t, x_t) v_t \right)^{-\frac{1}{\gamma}}. \tag{70}$$

Fixing x and v, the elasticity of this upper bound with respect to A is $(1 - \gamma - \eta_A(A,x))/\gamma$. Since this is here bounded below 0, (70) tends to 0 as $A \to \infty$. Finally, v_t is eventually positive, because by A4 A eventually exceeds 1 (rendering $v_t > 0$ feasible with $x = 1$ permanently), and v_t does not fall because sufficient precautions on new technology—e.g. banning its use—allow the consumption path to be maintained without increasing risk, by D4. Therefore, if $\eta_A(\cdot)$ is bounded above $1 - \gamma$, maintaining optimality condition (69) as $A_t \to \infty$ requires $x_t \to 0$.

A.5 Proof of Proposition 5

If $\lim_{k \downarrow 1} \tilde{R}(k) < 1$, there is a $\overline{k} > 1$ such that

$$\lim_{t\to\infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{\overline{k}}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{\overline{k}\gamma}{\gamma-1}}}{A_t \rho(\gamma-1)} < 1. \tag{71}$$

Choose $k \in (1, \overline{k})$. Suppose that $\nexists \underline{t} \ : \ C_t > t^{\frac{k}{\gamma-1}} \ \ \forall t > \underline{t}$. Then there is an increasing and unbounded sequence of times, $\{t_n\} \to \infty$, such that

$$C_{t_n} \leq t_n^{\frac{k}{\gamma-1}} \ \ \forall n \geq 1. \tag{72}$$

Observe that

$$\begin{aligned} &\lim_{n\to\infty} \frac{\partial \delta}{\partial x}\Big(A_{t_n}, \frac{t_n^{\frac{k}{\gamma-1}}}{A_{t_n}}\Big) \frac{t_n^{\frac{k\gamma}{\gamma-1}}}{A_{t_n} \rho(\gamma-1)} \\ \leq &\lim_{t\to\infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{\overline{k}}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{\overline{k}\gamma}{\gamma-1}}}{A_t \rho(\gamma-1)} \cdot t^{-\frac{\overline{k}-k}{\gamma-1}\gamma} = 0, \end{aligned} \tag{73}$$

where the inequality follows from the fact that, by D5, $\frac{\partial\delta}{\partial x}(A,x)$ weakly increases in x, and the limit before the $t^{-\frac{\overline{k}-k}{\gamma-1}\gamma}$ term is less than 1 by (71).

By (72), (73), and the fact that $v_t < \frac{1}{\rho(\gamma-1)}$ for all t, there is an $\underline{n}$ such that, for all $n \geq \underline{n}$,

$$\frac{\partial \delta}{\partial x}\Big(A_{t_n}, \frac{C_{t_n}}{A_{t_n}}\Big) v_{t_n} < A_{t_n} C_{t_n}^{-\gamma}.$$

This is compatible with optimality only if $x_{t_n} = A_{t_n} x_{t_n} = 1$. But this is impossible for sufficiently large n, by (29) and (72).

So for some $k > 1$,

$$\exists \underline{t} \ : \ C_t > t^{\frac{k}{\gamma-1}} \ \ \forall t > \underline{t}. \tag{74}$$

So (74) holds for $k = 1$ as well.

Given (74) for some $k > 1$, we have, for some $\underline{t}$ and some $\underline{k} \in (1, k)$, that for all $t > \underline{t}$

$$\begin{aligned} (A_t x_t)^{1-\gamma} &< t^{-k} \\ \implies \frac{\partial \delta}{\partial x}(A_t, x_t) x_t \, v_t &< t^{-k} \\ \implies \delta_t v_t &< t^{-k} \\ \implies \delta_t &< t^{-\underline{k}}. \end{aligned} \tag{75}$$

The first implication follows from the fact that $A_t^{1-\gamma} x_t^{-\gamma} \geq \frac{\partial \delta}{\partial x}(A_t, x_t) v_t$ whether or not x is interior. The second follows from the fact that $\delta < \frac{\partial \delta}{\partial x} x$ by D1 and D5. The third follows from the fact that v_t is eventually positive and does not fall to zero.

δ_t is uniformly bounded from 0 to $\underline{t}$ by $\max_{A \in [A_0, A_{\underline{t}}]} \delta(A, 1)$, which exists and is finite by the continuity of $\delta(\cdot)$ (D3). It follows from this and from (75) that $S_\infty > 0$.

If $\lim_{k \uparrow 1} \tilde{R}(k) > 1$, there is a $k < 1$ and an $\underline{s}$ such that

$$\frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{k}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{k\gamma}{\gamma-1}}}{A_t \rho(\gamma - 1)} > 1 \ \ \forall t > \underline{s}. \tag{76}$$

Suppose by contradiction that $\nexists \underline{t} \ : \ C_t < t^{\frac{1}{\gamma-1}} \ \ \forall t > \underline{t}$. Then there is an increasing and unbounded sequence of times, $\{t_n\} \to \infty$, such that

$$C_{t_n} \geq t_n^{\frac{1}{\gamma-1}} \ \ \forall n \geq 1. \tag{77}$$

Observe that

$$\begin{aligned} &\lim_{n\to\infty} \frac{\partial \delta}{\partial x}\Big(A_{t_n}, \frac{t_n^{\frac{1}{\gamma-1}}}{A_{t_n}}\Big) \frac{t_n^{\frac{\gamma}{\gamma-1}}}{A_{t_n} \rho(\gamma - 1)} \\ \geq &\lim_{t\to\infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{k}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{k\gamma}{\gamma-1}}}{A_t \rho(\gamma - 1)} \cdot t^{\frac{1-k}{\gamma-1}\gamma} = \infty, \end{aligned} \tag{78}$$

where the inequality follows from the fact that, by D5, $\frac{\partial\delta}{\partial x}(A,x)$ weakly increases in x, and the limit before the $t^{\frac{1-k}{\gamma-1}\gamma}$ term is greater than 1 by (76).

By (77), (78), and the fact that $v_t \not\to 0$, there is an n such that

$$\frac{\partial\delta}{\partial x}\Big(A_{t_n}, \frac{C_{t_n}}{A_{t_n}}\Big)v_{t_n} > A_{t_n}C_{t_n}^{-\gamma}.$$

This is incompatible with optimality. So

$$\exists \underline{t} : C_t < t^{\frac{1}{\gamma-1}} \;\; \forall t > \underline{t}. \tag{79}$$

By (79) and (29), $x_t \to 0$. So there exists a $\bar{t} \geq \underline{t}$ such that, for all $t > \bar{t}$, the choice of x is interior

$$\frac{\partial\delta}{\partial x}(A_t, x_t)v_t = A_t^{1-\gamma}x_t^{-\gamma}$$

and so, by (79),

$$\frac{\partial\delta}{\partial x}(A_t, x_t)x_t\, v_t = C_t^{1-\gamma} > 1/t.$$

Since $\eta_x \equiv \frac{\partial\delta}{\partial x}\frac{x}{\delta}$,

$$\eta_x(A_t, x_t)\delta(A_t, x_t)\, v_t > 1/t \;\; \forall t \geq \bar{t}.$$

Recall that an interior choice of x_t implies that $v_t > 0$, that v is upper-bounded by $\frac{1}{\rho(\gamma-1)}$, and that $\delta_t > 0$ by D1. So $\eta_x > 0 \; \forall t \geq \underline{t}$. So if η_x is upper-bounded by $\overline{\eta_x}$,

$$\delta(A_t, x_t) > \frac{\rho(\gamma-1)}{\overline{\eta_x}} \cdot \frac{1}{t} \;\; \forall t \geq \bar{t}.$$

So $S_\infty = 0$.

A.6 Proof of Proposition 6

Recalling that $\tilde{A}_0 = A_0$,

$$X\big(\tilde{A}(\cdot)\big) = \int_{A_0}^{\infty} \dot{\tilde{A}}_A^{-1}\, \delta\big(A, x_A[\tilde{A}(\cdot)]\big)dA,$$
$$X\big(A(\cdot)\big) = \int_{A_0}^{\infty} \dot{A}_A^{-1}\, \delta\big(A, x_A[A(\cdot)]\big)dA.$$

Since $\tilde{A}(\cdot)$ is an acceleration to $A(\cdot)$, there is a $\tau > 0$ such that $\dot{\tilde{A}}_A = \dot{A}_A$ and $x_A[A(\cdot)] = x_A[\tilde{A}(\cdot)]$ for all $A > \tilde{A}_\tau$.

If $X\big(A(\cdot)\big) < \infty$,

$$X\big(A(\cdot)\big) - X\big(\tilde{A}(\cdot)\big) = \int_{A_0}^{\tilde{A}_\tau} \Big(\dot{A}_A^{-1}\delta\big(A, x_A[A(\cdot)]\big) - \dot{\tilde{A}}_A^{-1}\delta\big(A, x_A[\tilde{A}(\cdot)]\big)\Big)dA \tag{80}$$

and $\dot{\tilde{A}}_A > \dot{A}_A$ for $A \in (A_0, \tilde{A}_\tau)$.

At any technology level A, any subsequent time-path of hazard rates feasible given $A(\cdot)$ is feasible with weakly higher consumption levels given $\tilde{A}(\cdot)$, by D5. So $v_A[\tilde{A}(\cdot)] \geq v_A[A(\cdot)]$ for all A. So, since

$$x_A[\tilde{A}(\cdot)] = \begin{cases} 1, & A^{1-\gamma} \geq \frac{\partial\delta}{\partial x}(A,1)v_A[\tilde{A}(\cdot)]; \\ A^{\frac{1-\gamma}{\gamma}}\left(\frac{\partial\delta}{\partial x}(A, x_A[\tilde{A}(\cdot)])v_A[\tilde{A}(\cdot)]\right)^{-\frac{1}{\gamma}}, & \text{otherwise,} \end{cases}$$

since $x_A[A(\cdot)]$ is defined likewise, and since $\frac{\partial\delta}{\partial x}(A,x)$ weakly decreases in x by D5, we have $x_A[\tilde{A}(\cdot)] \leq x_A[A(\cdot)]$ for $A \in (A_0, \tilde{A}_\tau)$ (indeed for all A).

It follows from D1, D2, and D5 that $\delta(\cdot)$ weakly increases in x. Thus $\dot{A}_A^{-1}\delta\big(A, x_A[A(\cdot)]\big) > \dot{\tilde{A}}_A^{-1}\delta\big(A, x_A[\tilde{A}(\cdot)]\big)$ for all $A \in (A_0, \tilde{A}_\tau)$, and (80) is positive.

If $\tau < \infty$, (80) finite. So if $X(A(\cdot)) = \infty$ and $\tau < \infty$, $X(\tilde{A}(\cdot)) = \infty$.

If $X(A(\cdot)) = \infty$ and $\tau = \infty$, it will suffice to find an example under which $X(\tilde{A}(\cdot))$ is finite and an example under which it is infinite. We have already encountered both.

For a case in which $X(\tilde{A}(\cdot)) = \infty$, consider the hazard function $\delta(A_t, x_t) = \bar{\delta}(A_t x_t)^\alpha$, discussed following Proposition 5. As discussed there, cumulative risk is then infinite for any technology path eventually bounded above zero.

For a case in which $X(\tilde{A}(\cdot)) < \infty$, consider hazard function (2)—$\delta(A_t, x_t) = \bar{\delta}A_t^\alpha x_t^\beta$—with $A(t) = t^k$, $\tilde{A}(t) = t^{\tilde{k}}$ for $t \geq 0$, where

$$k \leq \frac{\beta + \gamma - 1}{(\alpha - \beta)(\gamma - 1)} < \tilde{k}.$$

As explained in Section 4.2, under "Growth effects", $X(A(\cdot)) = \infty$ and $X(\tilde{A}(\cdot)) < \infty$.

A.7 Safety in redundancy, from discrete to continuous

Suppose a unit of production carries a constant flow probability $\bar{\delta}$ of triggering an existential catastrophe, so that, in the absence of any safeguards, the probability that it does not trigger a catastrophe after s units of time is $e^{-\bar{\delta}s}$. To be consistent with the discrete-time specification that the probability that it triggers a catastrophe after 1 unit of time equals p, we have $1 - e^{-\bar{\delta}} = p$ and thus $\bar{\delta} = -\log(1-p)$.

With $\frac{1-x_t}{x_t}$ units of safeguards maintained around t, since each unit multiplies the probability of a catastrophic failure per unit time by a factor $\tilde{b} \in (0,1)$, we have that the probability that a catastrophe is avoided until $t+s$ equals $e^{-\bar{\delta}\tilde{b}^{\frac{1-x}{x}}s}$.

The probability that $A_t x_t$ equally-safeguarded units of production all avoid catastrophe until $t+s$ is thus

$$\left(e^{-\bar{\delta}b^{\frac{1-x_t}{x_t}}s}\right)^{A_t x_t} = e^{-\bar{\delta}b^{\frac{1-x_t}{x_t}}A_t x_t s}. \tag{81}$$

So the probability of a catastrophe by s given locally constant A, x equals 1-(81), and the hazard rate—the probability of catastrophe per unit time—at time t precisely is

$$\delta_t \equiv \lim_{s \to 0} \left(1 - e^{-\bar{\delta}b^{\frac{1-x_t}{x_t}}A_t x_t s}\right)/s = \bar{\delta}A_t x_t \tilde{b}^{\frac{1-x_t}{x_t}}.$$

Letting $b \equiv -\log(\tilde{b}) > 0$ yields

$$\delta_t = \bar{\delta}A_t x_t e^{-b\frac{1-x_t}{x_t}}.$$

A.8 Proof of Proposition 7

By Appendix A.1, there is a unique optimal path. By the reasoning following (8), the optimal choice of x is 1 until the (unique) time at which

$$\frac{\partial u}{\partial x_t}(A_t, x_t) = \frac{\partial \delta}{\partial x_t}(A_t, x_t)\, v_t, \tag{82}$$

at $x_t = 1$, after which the optimal choice of x_t is interior and maintains equality (83).

Differentiating the utility function and hazard function (35), we have

$$
\begin{aligned}
A_t^{1-\gamma} x_t^{-\gamma} &= \bar{\delta} A_t e^{-b\frac{1-x_t}{x_t}} \Big(1 + \frac{b}{x_t}\Big) v_t \\
\implies \frac{1}{v_t} &= \bar{\delta} A_t^{\gamma} e^{-b\frac{1-x_t}{x_t}} \Big(x_t^{\gamma} + b x_t^{\gamma-1}\Big).
\end{aligned}
\tag{83}
$$

Because v_t increases monotonically and is upper-bounded, it is asymptotically positive and constant, by the monotone convergence theorem.

We must have $C_t \to \infty$. If we do not, then there is a unbounded sequence of times t_n and a consumption level $\overline{C}$ such that

$$
x_{t_n} \leq \overline{C}/A_{t_n} \;\; \forall n. \tag{84}
$$

Substituting (84) into (83), and recalling that $A_{t_n} \to \infty$, this would imply that the right-hand side of (83) tends to 0 across $\{t_n\}$, and thus that it is not asymptotically positive.

From (83),

$$
\frac{1}{v_t} = \delta_t C_t^{\gamma-1}(1 + b/x_t).
$$

Since $C_t^{\gamma-1} \to \infty$, x_t cannot be negative, and $1/v_t \not\to \infty$, it follows that $\delta_t \to 0$.

Since $C_t \to \infty$ and $\delta_t \to 0$, $v_t \to \frac{1}{\rho(\gamma-1)}$.

Divide both sides of (83) by $\bar{\delta} A_0^{\gamma}$, and take the log and then the limit. With

$$
\kappa \equiv \log\Big(A_0^{-\gamma} \frac{1}{\rho(\gamma-1)\bar{\delta}}\Big),
$$

we have

$$
\begin{aligned}
\lim_{t\to\infty} \Big[g\gamma t - b\frac{1-x_t}{x_t} + \log\Big(x_t^{\gamma} + b x_t^{\gamma-1}\Big)\Big] &= \kappa \\
\implies \lim_{t\to\infty} \frac{x_t}{1-x_t} t = \lim_{t\to\infty} &\frac{b}{g\gamma - \kappa/t + \log\big(x_t^{\gamma} + b x_t^{\gamma-1}\big)/t}.
\end{aligned}
$$

Other than $g\gamma$, the terms in the denominator on the right-hand side must converge to 0. This would be avoided only if there were an unbounded sequence of times t_n across which x_{t_n} grew at least exponentially with time, which is impossible, or shrank at least exponentially with time, which would send the right-hand side of (83) to zero. So

$$\begin{aligned}
&\lim_{t\to\infty} \frac{x_t}{1-x_t} t = \frac{b}{g\gamma} \\
\implies &\lim_{t\to\infty} x_t t = \lim_{t\to\infty} (1-x_t)\frac{b}{g\gamma} = \frac{b}{g\gamma} \\
\implies &\lim_{t\to\infty} x_t \frac{g\gamma}{b} t = 1,
\end{aligned}$$

since $x_t \to 0$. It then follows from the hazard function that, in the limit, δ falls to 0 at exponential rate $-g(\gamma-1) < 0$.

B Transition dynamics for simulations

For simulating the transition dynamics, it is helpful to find $\dot{x}_t$ and $\dot{\delta}_t$ as functions of t and x_t in the regime where x is interior.

Hazard function (2), used throughout Sections 2–4 and used to simulate Figures 1 and 2, is the special case of hazard function (33), used to simulate Figure 3, with $\epsilon = 1$. The calculations below therefore apply to all simulations.

FOC:

$$\begin{aligned}
&\frac{\partial u}{\partial x_t}(A_t, x_t) = \frac{\partial \delta}{\partial x_t}(A_t, x_t) v_t \\
\implies &A_t^{1-\gamma} x_t^{-\gamma} = \bar{\delta} A_t^{\alpha} x_t^{\beta-2}\Big((\beta-1)\big(1-(1-x_t)^{\epsilon}\big) + \epsilon x_t (1-x_t)^{\epsilon-1}\Big) v_t.
\end{aligned}$$

Rearranging and differentiating gives

$$v_t = \frac{1}{\bar{\delta}} \frac{A_t^{1-\gamma-\alpha} x_t^{2-\gamma-\beta}}{(\beta-1)\big(1-(1-x_t)^{\epsilon}\big) + \epsilon x_t (1-x_t)^{\epsilon-1}} \tag{85}$$

$$\begin{aligned}
\implies \dot{v}_t = v_t\Big(&(1-\gamma-\alpha)g + (2-\gamma-\beta)\frac{\dot{x}_t}{x_t} \\
&- \epsilon \frac{\beta - (\epsilon+\beta-1)x_t}{(\beta-1)(1-x_t)^{1-\epsilon} + 1 - \beta + (\epsilon+\beta-1)x_t} \frac{\dot{x}_t}{1-x_t}\Big).
\end{aligned} \tag{86}$$

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

$$\begin{aligned}\dot{v}_t &= v_t(\rho + \delta_t) - u(c_t) \\ &= v_t\Big(\rho + \bar{\delta} A_t^{\alpha} x_t^{\beta-1}(1-(1-x_t)^{\epsilon})\Big) - \frac{(A_t x_t)^{1-\gamma} - 1}{1-\gamma}. \end{aligned} \tag{87}$$

Substituting (85) into (86) and (87), setting the results equal, and solving for $\dot{x}_t$ yields

$$\begin{aligned}\dot{x}_t = x_t&\big((\beta-1)(1-x_t)^{1-\epsilon} + 1 - \beta + (\epsilon+\beta-1)x_t\big)(1-x_t) \\ &\Big((2-\gamma-\beta)\big((\beta-1)(1-x_t)^{1-\epsilon} + 1 - \beta \\ &+ (\epsilon+\beta-1)x_t\big)(1-x_t) - \epsilon(\beta-(\epsilon+\beta-1)x_t)x_t\Big)^{-1} \\ &\Big(\rho + \bar{\delta}A_t^{\alpha}x_t^{\beta-1}(1-(1-x_t)^{\epsilon}) - g(1-\alpha-\gamma) - \\ &\frac{(A_t x_t)^{1-\gamma}-1}{1-\gamma}\bar{\delta}A_t^{\alpha+\gamma-1}x_t^{\beta+\gamma-2}\big((\beta-1)(1-(1-x_t)^{\epsilon}) + \epsilon x_t(1-x_t)^{\epsilon-1}\big)\Big).\end{aligned} \tag{88}$$

Differentiating the hazard function (33) with respect to t yields

$$\dot{\delta}_t = \bar{\delta}A_t^{\alpha}x_t^{\beta}\frac{1-(1-x_t)^{\epsilon}}{x_t}\Big(\alpha g + (\beta-1)\frac{\dot{x}_t}{x_t} + \epsilon\frac{(1-x_t)^{\epsilon}}{1-(1-x_t)^{\epsilon}}\frac{\dot{x}_t}{1-x_t}\Big). \tag{89}$$

Scripts for replicating Figures 1, 2, and 3 using (88) and (89), and the estimate of S_∞ following Figure 1, are provided here: https://philiptrammell.com/static/ERAG_code.zip.