Existential Risk and Growth

Philip Trammell* and Leopold Aschenbrenner†

December 7, 2024

Technologies may pose existential risks to civilization. Though accelerating technological development may increase the risk of anthropogenic existential catastrophe per period in the short run, two considerations suggest that a sector-neutral acceleration decreases the risk that such a catastrophe *ever* occurs. First, acceleration decreases the time spent at each technology level. Second, since a richer society is willing to sacrifice more for safety, optimal policy can yield an "existential risk Kuznets curve"; acceleration then pulls forward a future in which risk is low. Acceleration typically increases risk only given sufficiently extreme policy failures or direct contributions of acceleration to risk. *JEL codes: O32, O33*

*Global Priorities Institute and Department of Economics, University of Oxford. Contact: philip.trammell@economics.ox.ac.uk. We thank Danny Bressler, Lennart Stern, and Michael Wiebe for suggesting the idea; Ben Snodin, Luis Mota, Tyler Cowen, Rick van der Ploeg, Pete Klenow, Chad Jones, Toby Ord, and attendees of several workshops and work-in-progress seminars at the Global Priorities Institute for helpful comments; and Alex Holness-Tofts and Arvo Muñoz for research assistance on an earlier draft. This paper was written with support from Open Philanthropy, the Future of Humanity Institute (University of Oxford), and the Centre for Effective Altruism.

†Situational Awareness LP and Global Priorities Institute, University of Oxford.

1 Introduction

Technology brings prosperity. On the other hand, some technological developments have arguably raised, or would raise, *existential risk*: the risk of human extinction or of an equally complete and permanent loss of human welfare.[1]

This raises a possible tradeoff: concern for the survival of civilization may motivate slowing development. Environmentalist sentiments along these lines go back at least to the Club of Rome's 1972 report on the "Limits to Growth", and have arguably reemerged with calls to pause AI development (Future of Life Institute, 2023). Jones (2024) explores how to trade off between AI development and AI risk, assuming the tradeoff exists.

Even if some technological developments directly raise existential risk, however, others may directly lower it. Advances in game theory may render us less vulnerable to nuclear war; vaccines render us less vulnerable to plagues. The prosperity technology brings can also lower existential risk indirectly, by increasing a planner's willingness to pay for safety. This paper offers an argument that these salutary possibilities probably dominate in the long run, and that the proposed tradeoff is thus typically illusory. That is, concern for long-term survival should typically motivate speeding rather than slowing technological development.

We begin in Section 2 with a simple model in which the hazard rate—the probability of catastrophe per period—is a positive function of the technology level. Here, an existential catastrophe must occur unless higher technology levels carry hazard rates that eventually fall toward zero.

In this setting there are only two possibilities. If advanced technology does not eventually drive the hazard rate toward zero, then a catastrophe is inevitable, so accelerating technological development cannot increase its probability. Otherwise, a catastrophe is avoidable, and acceleration can lower its probability by hastening the arrival of safety.

This simple model formalizes two observations. First, if we believe the hazard rate is currently high, our only hope for a long future is the hope that we are in a temporary "time of perils". This view was famously expressed by Sagan (1997), and its implications for those especially concerned about the long-term future are emphasized by Parfit (1984), Ord (2020), and others. The second observation is less widely appreciated: that if we are in a time of perils, with the hazard rate a positive function only of the technology level, then deceleration for the sake of long-term survival is misguided. Speeding technological development may be temporarily risky,

[1]See e.g. Bostrom (2002), Posner (2004), Farquhar et al. (2017), Ord (2020), and Jones (2024). We will refer to the event that humanity immediately goes extinct or suffers a similarly complete and permanent loss of welfare as an "existential catastrophe", or simply "catastrophe". We will refer to "humanity" and "human civilization" interchangeably.

but in this setting it must be safer in the long run.[2]

The model of Section 2 is not "economic". It studies the impact on risk of quickly escaping risky states, not optimal policy under constraints. It thus leaves open the possibility that, when consumption–risk tradeoffs are navigated by a planner with little concern for long-term survival, technological acceleration can increase risk after all. Section 2 also offers no reason to believe that future states *will* be safe. If one believes that technology has historically increased the hazard rate, the hope that this relationship will reverse in the future may seem naive. As Thorstad (2022) emphasizes, the "time of perils hypothesis" has to date largely been asserted without strong defense.[3]

Section 3 therefore introduces an environment in which technology grows exogenously and its risks can be mitigated by policy. As dangerous new technologies are introduced, a planner, discounting the future at an arbitrary rate, decides how much consumption to sacrifice to lower the hazard rate. We illustrate that, even if technological advances always directly raise the hazard rate, optimal policy can generate an "existential risk Kuznets curve", with the hazard rate rising and then falling as technology advances. Early, when the expected discounted value of the future of civilization is low and the marginal utility of consumption is high, it is worthwhile to adopt risky technologies as they arrive. Later, when the discounted future is more valuable and the marginal utility of consumption has fallen, substantial risk mitigation becomes worthwhile.

The possibility of rational policy thus offers an economic justification for the view that we may indeed be living through a once-in-history time of perils. Safety is a luxury. Technological development generates a wealth effect. If the utility function is concave enough—i.e. the wealth effect strong enough—optimal policy lowers the hazard rate quickly enough that the probability of long-term survival is positive. 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, 2024), on which growth increases the value of life relative to marginal consumption. Like the analysis presented here, these papers find that, given a concave enough utility function, enrichment motivates large reallocations from consumption to safety. None of these sources solve for the optimal path of a hazard rate over time, however, or characterize conditions under which the probability of a binary event (here, existential catastrophe) under optimal policy is less than 1.[4]

[2]The point is however noted informally by Bostrom (2014), p. 234, and recently by Ord (2024).

[3]As of the time of this writing, Thorstad cites an earlier draft of this paper as an example of an argument for the hypothesis, but criticizes the model offered in that draft. We believe the model in the current paper is robust to this criticism.

[4]Our model of catastrophic risk differs more significantly 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 equivalent to, proportional consumption cuts. In such a context there are no wealth effects: the fraction of consumption one is willing to sacrifice to

Relative to the model of Section 2, optimal policy tends to magnify the extent to which technological acceleration decreases long-term risk, for two reasons:

1. As in the policy-free model, acceleration decreases the time spent in any given risky state. Under optimal policy, however, the wealthier future states pulled forward by an acceleration are systematically inclined to be safer, due to the wealth effect.

2. Given an increase to the *future* growth rate, even before actual productive capacity has yet increased, the anticipation of a more valuable future motivates more stringent safety policy in the present.

Sections 2 and 3 explore models in which the *state* of technology at a given time contributes to the hazard rate. Section 4 considers the possibility that risk is "transitional", increasing in the *rate of technological development*.

Absent policy, the effect of acceleration on long-term transition risk is ambiguous. In particular, acceleration has no effect on long-term risk if the "experiment" associated with developing a given technology poses a risk independent of how many experiments happen concurrently. This is the assumption of e.g. Jones's (2016) "Russian roulette" model of risky technological development. If the future contains a sequence of experiments, each of which will pose some risk of catastrophe, then stagnation can lower risk by avoiding advanced experiments altogether; but an acceleration that only pulls forward their date leaves the probability of catastrophe unchanged.

As in Section 3, introducing an optimal policy response facilitates survival due to wealth effects, potentially replacing an ever-increasing hazard rate with a Kuznets curve. Also, though the effect of acceleration on long-term transition risk remains ambiguous given policy, policy can shift the conditions under which acceleration has a given effect on risk. At least in the particular model of transition risk studied in Section 4, the existence of a policy response significantly widens the conditions under which acceleration lowers transition risk.

Section 5 summarizes these analyses and their limitations.

---

avoid a proportional consumption cut is, by definition, independent of one's consumption.

This analysis might better be compared with that of Baranzini and Bourguinion (1995). Baranzini and Bourguinion find conditions under which the growth path that maximizes expected discounted utility also minimizes the probability of existential catastrophe. In our model these objectives never perfectly align, but we explore how technological advances, when regulated with a view to maximizing expected discounted utility, can lower the probability of existential catastrophe.

2 State risk without policy

2.1 Model

The hazard rate — The "hazard rate" δ_t is the flow probability at t of anthropogenic existential catastrophe. In this section we posit that it is a positive, continuous function of a state variable A_t:

$$\delta_t = \delta(A_t), \quad \delta(A) > 0 \ \ \forall A.$$

Assume that $A_{(.)}$ is differentiable, with a positive derivative, and that it increases without bound.

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

Survival — The probability that civilization survives to date t is given by

$$S_t \equiv e^{-\int_0^t \delta_s ds} \quad \Longleftrightarrow \quad \dot{S}_t = -\delta_t S_t, \ \ S_0 = 1.$$

The probability that human civilization avoids a catastrophe and, at least in expectation, enjoys a long and flourishing future[5] is

$$S_\infty \equiv \lim_{t\to\infty} S_t = e^{-\int_0^\infty \delta_s ds}. \tag{1}$$

We will refer to $\{\delta_t\}_{t=0}^\infty$ as the *hazard curve*, to the area under the hazard curve $\int_0^\infty \delta_t dt$ as *cumulative risk*, and to S_∞ as the *probability of survival*.

Note that the probability of survival decreases in cumulative risk, and survival is possible ($S_\infty > 0$) iff cumulative risk is finite. Existential catastrophe is not guaranteed only if, roughly, the world is on track eventually to grow ever safer.

[5] In the face of natural existential risk, this will entail 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 below one in 870,000 per year. Throughout this paper we ignore the possibility that technological advances may mitigate natural existential risks. Accounting for this possibility would only strengthen the headline results.

2.2 How does acceleration affect risk?

We will now consider how accelerating the overall technology path affects cumulative risk. Importantly, in practice, we may often be able to accelerate or delay particular technologies in isolation. Supporting vaccine technologies while discouraging technologies that facilitate engineering pathogens may well lower existential risk from pandemics more than speeding or slowing research across the board. **Such considerations are discussed in Section 3.1.** Here, however, we begin by fixing a path and studying the implications of faster movement along it.

Absent a negative shock severe enough to induce stagnation or recession, technology crosses every value from A_0 to ∞ exactly once. So the area under the hazard curve can be found by integrating with respect to technology instead of time:

$$\int_0^\infty \delta(A_t)dt = \int_{A_0}^\infty \delta(A)\Big(\frac{dA}{dt}\Big)^{-1}dA = \int_{A_0}^\infty \delta(A)\dot{A}_A^{-1}dA, \tag{2}$$

where, somewhat abusing notation, $\dot{A}_A$ denotes the value of $\dot{A}$ when the technology level equals the subscripted A. This change of variables makes it easier to see how various shocks to the growth path affect cumulative risk.

Instantaneous level effects — Consider a shock to the technology level for a short period beginning at t, so that the technology level over this period is approximately $\tilde{A}$ rather than A_t (and the subsequent technology path is unchanged). The sign of the impact of this shock on cumulative risk depends on whether $\delta(\tilde{A})$ is greater or less than $\delta(A_t)$. By the leftmost integral of (2), the impact on cumulative risk *per unit time* of this shock to the technology level at t equals

$$\delta(\tilde{A}) - \delta(A_t).$$

Instantaneous accelerations — Consider the impact on cumulative risk per unit time of an instantaneous shock to technology *growth* at t, so that the technology growth rate at t is $\dot{\tilde{A}}$ rather than $\dot{A}_t$, and the subsequent technology growth rate at each technology level is unchanged. By the rightmost integral of (2), the impact of this shock on cumulative risk *per unit of increase to the technology level* during the acceleration is $\delta(A_t)(\dot{\tilde{A}}^{-1} - \dot{A}_t^{-1})$. Multiplying this by the new rate of technology growth per unit time, the impact on cumulative risk per unit time equals

$$-\delta(A_t)\big(\dot{\tilde{A}}/\dot{A}_t - 1\big).$$

Accelerations — Choose technology level $\overline{A} > A_t$. Since the baseline technology path increases continuously and without bound, $\overline{A} = A_T$ for some $T > t$.

Consider the effect of increasing the technology level at t from A_t to $\overline{A}$ and subsequently maintaining the technology path $\overline{A}_s = A_{s+(T-t)}$ $(s \geq t)$. This shock to the technology path amounts to a "leap forward in time", cutting a slice cut out of the hazard curve. Cumulative risk falls by

$$\int_{A_t}^{\overline{A}} \delta(A)\dot{A}_A^{-1}\, dA.$$

More generally, define a *temporary acceleration* as an increase to $\dot{A}$ at some range of technology levels: say, from A_t to A_T. Because the exponent on $\dot{A}$ in the integral is negative, the 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 A_T.

A jump in the technology level from A_t to A_T temporarily increases the hazard rate if $\delta(A_T) > \delta(A_t)$. Likewise, an acceleration to technology growth accelerates an increase to the hazard rate if $\delta(\cdot)$ is increasing around A_t. It may therefore appear to contemporaries that a given permanent level effect decreases the probability of survival. Here, that would be incorrect. If (2) is finite, the permanent level effect decreases cumulative risk and increases the probability of survival. If (2) is infinite, the probability of survival is zero with or without the permanent level effect.[6] The two possibilities are illustrated below.

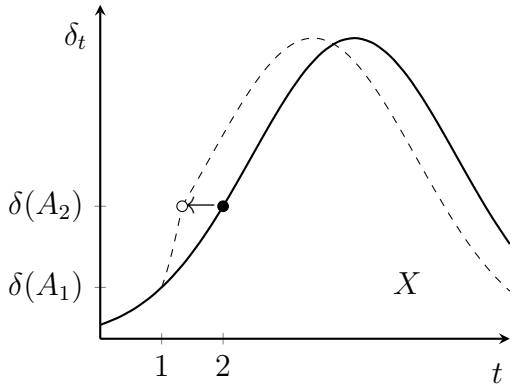


Figure 1a: $X < \infty$;
temp. acceleration lowers X

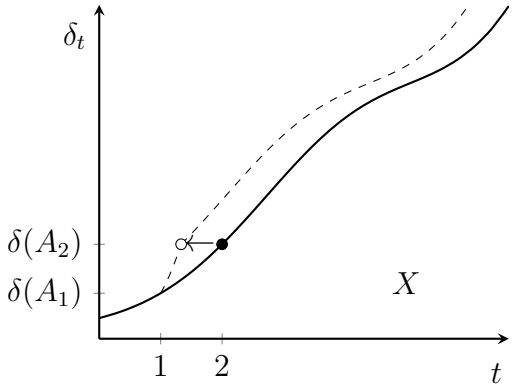


Figure 1b: $X = \infty$;
temp. accel. has no effect on X

Define a *permanent acceleration* to be a permanent increase to $\dot{A}$ from some time t—or, equivalently, some technology level A_t—onward. By the rightmost integral of (2), a permanent acceleration, like a temporary acceleration, must lower cumulative risk if cumulative risk is finite on the baseline technology path.

[6] $\int_{A_t}^{A_T} \delta(A_t)dt$ is finite by the continuity of δ in A and of A in t.

Unlike temporary accelerations, however, permanent accelerations can render survival possible when it would otherwise be impossible. Shrinking a heavy-tailed curve with an infinite integral can yield a thin-tailed curve with a finite integral.

To state this lesson in reverse, consider stagnation: a permanent negative acceleration, or "deceleration", setting $A_s = A_t$ for all $s \geq t$. The hazard rate is then permanently positive, and survival is impossible, even if it might have been possible with technology growth. More concretely, consider the implications of a large negative shock today returning the world to the state it inhabited in 1924. Perhaps the hazard rate was much lower in 1924 than today, but this reset would largely doom us to relive the nuclear standoffs, emissions-intensive industrializations, and biotechnological hazards of the past. With enough replays of the past century, a catastrophe would presumably be inevitable.

3 State risk with policy

3.1 Motivation

The previous section shows that, under weak conditions, acceleration along a "technology path" either lowers or does not affect cumulative risk. Any optimism one might draw from this observation about the risk implications of accelerating technological development in practice faces two limitations.

First, cumulative risk falls only if we are already on track to eventually grow ever safer. Because the technology path was assumed to be fully exogenous, the model of the previous section offers no reason to believe we are. If technological progress, broadly construed, has historically increased the hazard rate, the message of Section 2 is that those who wish to reduce existential risk should accelerate such progress on the blind hope that the associated risk trend eventually reverses.

Second, the result only applies to a perfectly balanced acceleration in the path of every determinant of the hazard rate. It says nothing about the implications of pulling forward some technologies in time without equally pulling forward all others.

To illustrate the second limitation, suppose the hazard rate is a function of two variables, A and x, and suppose the path of x is fixed:

$$\delta_t = A_t x_t, \quad x_t = (1+t)^{-2}.$$

Here, later technology states A_t are always riskier on balance. x_t may represent an index of policies and tools which promote safety ever more effectively with time and are independent the state of (acceleratable) technology. Consider an acceleration in the technology path from $A_t = (1+t)^k$ to $A_t = (1+t)^{\tilde{k}}$, where $k < 1 < \tilde{k}$. This

acceleration increases cumulative risk from

$$\int_0^\infty (1+t)^{k-2}dt \quad \text{to} \quad \int_0^\infty (1+t)^{\tilde{k}-2}dt.$$

The former is finite, because $k-2<-1$. The latter is infinite, because $\tilde{k}-2>-1$. In this case, accelerating A lowers the probability of survival to zero.

At the other extreme, however, suppose that the index x_t of risk-relevant features of the world not included in a given technological acceleration is set by policy, in light of the path of A, to optimize a tradeoff between safety and consumption. As long as more advanced technology levels make greater *consumption* feasible, we will see that optimal policy—with respect to any discount rate—generally *strengthens* the conclusion that technological acceleration lowers cumulative risk.

As in the tech-only model of Section 2, survival can only be achieved by pulling forward a future that asymptotically approaches perfect safety. Whereas the earlier model is agnostic about whether more advanced technology will in fact carry a lower hazard rate, however, an optimal policy response introduces a tendency for faster technological development to carry lower risk in the long run. This is because technology increases consumption, which both decreases the utility cost of a marginal consumption sacrifice and increases the value of life. Furthermore, the prospect of a *future* acceleration now lowers the *present* hazard rate, because when the value of the future is greater, it is worth sacrificing more today to prevent its ruin.

With reference to Figure 1: The first implication of optimal policy is that the finite X case is more likely. The second is that the hazard rate now decreases in anticipated future growth.

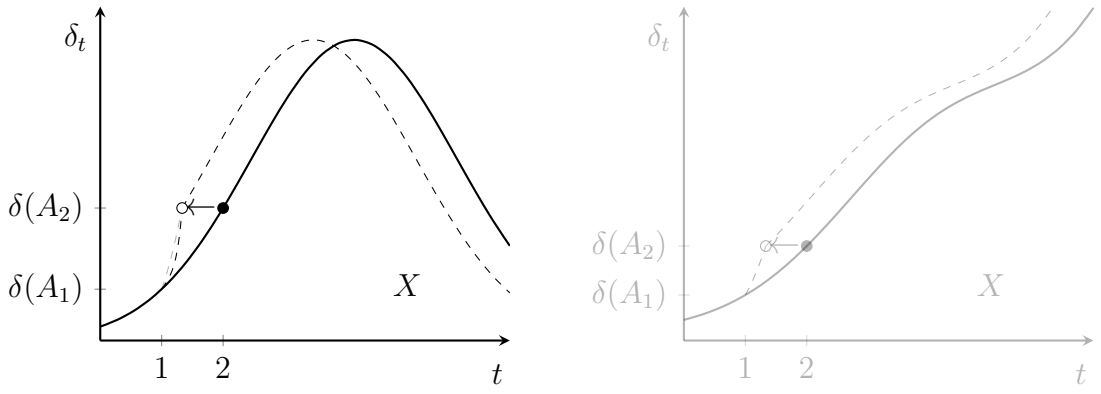


Figure 2: Optimal policy (i) facilitates finite X and
(ii) lowers the hazard rate associated with each technology
level while an acceleration is underway

These dynamics are illustrated in a simple model of technology and optimal policy

in the rest of this section. **Generalized results, not relying on functional form assumptions, are given in Appendix B.3.**

3.2 The economic environment

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

$$C_t = A_t x_t. \tag{3}$$

The tradeoff at the heart of this section is that a technologically advanced civilization can risk self-destruction, but 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 risky production processes and/or resource-allocations to safety-increasing services like pandemic detection.

The technology frontier A grows at a constant rate g:

$$\dot{A}_t = A_t g, \quad g > 0, A_0 > 1. \tag{4}$$

The hazard rate — The hazard rate δ_t is now a function of the technology level A_t and the policy choice $x_t \in [0, 1]$, and is increasing in x_t. In this illustrative model, the elasticities of the hazard rate in A and in x are constant:

$$\delta(A_t, x_t) = \bar{\delta} A_t^{\alpha} x_t^{\beta}, \quad \bar{\delta} > 0, \ \beta > \alpha > 0, \ \beta > 1. \tag{5}$$

We impose the three inequalities of $\beta > \alpha > 0$, $\beta > 1$ to satisfy three desiderata.[7]

The first is that, fixing $x_t > 0$, δ_t increase in A_t. This imposes $\alpha > 0$. The assumption that δ_t increase in A_t is necessary if we are to concede that technological development has historically increased the hazard rate, and that this trend would continue absent a change in policy.[8] It is thus necessary to illustrate that optimal policy can render survival possible when it might otherwise have been impossible.

[7] Hazard function (5) 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.

[8] The proportion $1 - x$ of potential consumption sacrificed for the sake of existential safety has only increased as technology has advanced. Ord (2020, p. 313) estimates that, as of 2020, approximately $100M/year was spent specifically on reducing existential risk. This is likely a great underestimate of existential safety expenditures in the sense relevant here, for two reasons. First, explicit expenditures do not include foregone consumption due to regulatory barriers. Second, many catastrophic risk reduction efforts are motivated both by the desire to reduce existential risks and by the desire to reduce smaller-scale damages. By contrast, Moynihan (2020) argues that the very concept of an anthropogenic existential catastrophe essentially did not exist 300 years ago; it appears there were then no efforts taken to prevent one.

Second, the elasticity of δ_t with respect to x_t is assumed to exceed its elasticity with respect to A_t; i.e. $\beta > \alpha$. This is equivalent to the condition that, when technology advances, it is feasible to lower the hazard rate by retaining the former consumption level, allocating all marginal productive capacity to safety measures. This may be seen by substituting $x_t = C_t/A_t$ (from (3)) into the hazard function (5), 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 model if $\beta \leq \alpha$, then an existential catastrophe is unavoidable unless consumption falls to zero. This degrowth would amount to the destruction of civilization by other means. If $\beta \leq \alpha$, therefore, speeding or slowing growth can have no impact on the probability of an existential catastrophe broadly construed.

Third, fixing $A_t > 0$, $\beta > 1$ renders δ_t strictly convex in x_t, so that there are diminishing returns to safety efforts. We consider this assumption reasonable both from first principles and from Shulman and Thornley's (2024) estimates of the cost-effectiveness of existential risk mitigation efforts (Appendix B.1).

Preferences — A planner maximizes expected discounted flow utility, which is a CRRA function of consumption:

$$\int_0^{\infty} e^{-\rho t} S_t \, u(C_t) \, dt; \qquad u(C_t) = \begin{cases} \frac{C_t^{1-\gamma}-1}{1-\gamma}, & \gamma > 0, \neq 1; \\ \log(C_t), & \gamma = 1. \end{cases} \tag{6}$$

The discount rate $\rho > 0$ is the sum of some rate of pure time preference and/or some rate of natural and unavoidable existential risk.[9] When $\gamma < 1$, we impose

$$\rho > \underline{\rho} \equiv \frac{(\beta-\alpha)(1-\gamma)}{\beta} g \tag{7}$$

to ensure the existence of an optimal policy path.

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.

Note that when $\gamma > 1$, flow utility is upper-bounded by $\frac{1}{\gamma-1}$. Accelerating consumption growth, from a baseline of positive consumption growth, therefore yields a stream of utility increases that eventually shrinks over time. Concern for the future then clearly casts doubt on the value of speeding technological development: the consumption benefits of doing so primarily accrue in the short run, whereas the costs

[9] One valid interpretation of these preferences is that the population is fixed and (6) is the expected utility of a representative household. Another is that population grows exponentially at rate $n < \rho$, that the rate of pure time preference and exogenous risk is in fact $\rho + n$, and that the planner uses the total utilitarian social welfare function.

of an existential catastrophe are everlasting. By contrast, when $\gamma \leq 1$, flow utility can grow without bound, so accelerations to consumption growth and reductions in existential risk can have comparable long-term benefits.

3.3 The existential risk Kuznets curve

Optimality — Summarizing Section 3.2, the planner chooses a policy path $\{x_t\}_{t=0}^{\infty}$ to maximize expected utility (6) subject to

$$\begin{aligned} &A_0 > 1, \;\; \dot{A}_t = gA_t \;\; (g > 0), \\ &C_t = A_t x_t, \\ &S_0 = 1, \;\; \dot{S}_t = -\delta_t S_t, \\ &\delta_t = \bar{\delta} A_t^{\alpha} x_t^{\beta} \;\; (\bar{\delta} > 0, \; \beta > \alpha > 0, \; \beta > 1). \end{aligned} \tag{8}$$

This section finds the path of the hazard rate in the planner's solution. The next section explores what this implies for the impact of acceleration on cumulative risk.

The planner faces one choice variable, x_t, and one state variable, S_t. Her expected flow utility at t is $S_t u(C_t)$. Her problem is represented by current-value Lagrangian

$$\begin{aligned} \mathcal{L}_t &= S_t u(C_t) + v_t \dot{S}_t + \mu_t (1 - x_t) \\ &= S_t \frac{(A_t x_t)^{1-\gamma} - 1}{1-\gamma} - v_t \, \bar{\delta} A_t^{\alpha} x_t^{\beta} S_t + \mu_t (1 - x_t). \end{aligned} \tag{9}$$

μ_t is the Lagrange multiplier on x, positive iff the $x_t \leq 1$ constraint binds.

$$v_t = \int_t^{\infty} e^{-\rho(s-t)} \frac{S_s}{S_t} u(C_s) ds \tag{10}$$

is the costate variable on survival: the expected value of civilization as of t.[10]

On an optimal path, the first-order condition on (9) with respect to the choice variable x_t is satisfied. Differentiating (9) with respect to x_t, we have

$$S_t A_t^{1-\gamma} x_t^{-\gamma} - \bar{\delta} A_t^{\alpha} \beta x_t^{\beta-1} \, v_t S_t \geq 0, \tag{11}$$

with inequality iff the left-hand side is positive at $x_t = 1$, in which case $x_t = 1$ is optimal.[11] So as long as (11) is nonnegative at $x_t = 1$, the optimal $x_t \in [0, 1]$ equals 1. Any consumption sacrifices would carry greater flow costs than expected benefits. When (11) is negative at $x_t = 1$, the optimal choice of x_t is interior. It sets (11) equal to zero, maintaining that on the margin, the loss of flow utility from lowering consumption equals the expected benefit via risk reduction.[12]

[10] The costate variable on survival equals (10) because the value of saving the world equals the expected value of the world. The equation is derived formally in Appendix A.1.

[11] The second derivative with respect to x_t is negative because $\beta > 1$.

[12] We can ignore the possibility that optimal x_t equals 0 because this yields infinite flow disutility.

In fact there is a unique[13] optimal path, characterized by (11), a first-order condition corresponding to S_t, and identity (10) (see Appendix A.1).

Initial risk increases — (11) is nonnegative at $x_t = 1$ iff

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

The continuation value of civilization at t given survival to t, v_t, strictly rises over time. This is because, given the best 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, given $\beta > \alpha$. A higher initial technology level makes a preferred future feasible.

Suppose inequality (12) is satisfied strictly at $t = 0$. Then early in time, when A_t is low, the optimal policy choice is $x = 1$, and the hazard rate rises at rate[14]

$$g_{\delta t} = \alpha g.$$

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 (11) equal to zero.

Setting (11) equal to zero and rearranging, we have the optimal choice of x_t after t^*, and thus the optimal choice of x_t in general:

$$x_t = \begin{cases} 1, & t \leq t^*; \\ \left(\bar{\delta}\beta A_t^{\alpha+\gamma-1} v_t\right)^{-\frac{1}{\beta+\gamma-1}}, & t > t^*. \end{cases} \tag{13}$$

Taking growth rates, we find the growth rate of the policy variable after t^*:

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

The hazard rate in turn grows as

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

Finally, on an optimal path, v grows asymptotically at a constant rate, with

$$\lim_{t\to\infty} g_{vt} = \begin{cases} \frac{(\beta-\alpha)(1-\gamma)}{\beta} g, & \gamma < 1; \\ 0, & \gamma \geq 1. \end{cases} \tag{16}$$

[13] Given semi-continuity. If path x is optimal, measure-zero deviations from x are also optimal.

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

This is proved in Appendix A.2, but an intuition is as follows. When $\gamma > 1$, growth in v must fall to zero because v_t is upper-bounded by

$$\bar{v} \equiv \frac{1}{\rho(\gamma - 1)}. \tag{17}$$

When $\gamma < 1$, flow utility grows approximately like $C_t^{1-\gamma}$ when C_t is large. Observe from (10) that v_t grows roughly with flow utility. Substituting $g_v = (1-\gamma)g_C = (1-\gamma)(g + g_x)$ into (15) then yields the $\gamma < 1$ case of (16).

Substituting (16) into (14) and (15) gives the asymptotic growth rates g_x and g_δ. Also, since $C_t = A_t x_t$, we can easily find the asymptotic value of g_C, which is always positive. Though x falls to 0, A grows more quickly than x declines. Indeed, as elaborated in Appendix B.3.3, consumption growth is key to the growth in sacrifices for safety. With decreasing marginal utility to *consumption* and decreasing marginal returns to *sacrifices for safety*, potential consumption is split between the former and latter so that the marginal value of each stays equal.

Proposition 1. ***The existential risk Kuznets curve***
On the path defined by (6)–(8), there is a time $t^* \geq 0$ such that for $t < t^*$,

$$x_t = 1, \quad g_{Ct} = g > 0, \quad g_{\delta t} = \alpha g > 0;$$

and for $t > t^*$ x_t is interior, such that if $\gamma > 1$,

$$\lim_{t\to\infty} g_{xt} = -\frac{\alpha + \gamma - 1}{\beta + \gamma - 1} g < 0, \tag{18}$$

$$\lim_{t\to\infty} g_{Ct} = \frac{\beta - \alpha}{\beta + \gamma - 1} g > 0, \tag{19}$$

$$\lim_{t\to\infty} g_{\delta t} = -\frac{(\beta - \alpha)(\gamma - 1)}{\beta + \gamma - 1} g < 0; \tag{20}$$

and if $\gamma \leq 1$,

$$\lim_{t\to\infty} g_{xt} = -\frac{\alpha}{\beta} g < 0, \tag{21}$$

$$\lim_{t\to\infty} g_{Ct} = \frac{\beta - \alpha}{\beta} g > 0, \tag{22}$$

$$\lim_{t\to\infty} \delta_t t = \frac{\rho}{(\beta - \alpha) g} > 0, \quad \gamma = 1; \tag{23}$$

$$\delta^* \equiv \lim_{t\to\infty} \delta_t = \frac{(\rho - \underline{\rho})(1 - \gamma)}{\beta + \gamma - 1} > 0, \quad \gamma < 1. \tag{24}$$

Proof. See Appendix A.2. □

Corollary 1.1. ***Survival***
$S_\infty > 0$ iff $\gamma > 1$.

Proof. The "if" follows from (20) and the definition of S_∞. δ_t ultimately falls exponentially, so $\int_0^\infty \delta_t dt < \infty$, so $S_\infty \equiv e^{-\int_0^\infty \delta_t dt} > 0$. The "only if" likewise follows from (23)–(24): when the hazard rate is asymptotically constant, or falls proportionally to $1/t$ (or more slowly), the integral of the hazard curve diverges. □

Intuition for the $\gamma = 1$ threshold — The importance of γ stems from the fact that, 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 the hazard rate × the value of civilization.

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 $1/t$.

This result recalls the "Russian roulette" model of Jones (2016). There, optimal policy renders catastrophe avoidable iff $\gamma \geq 1$. Since risk in that model is posed by the development (not existence) of technologies, it more similar to the "transition risk" model of Section 4, and is discussed further there.

Simulation — The paths of policy and risk are simulated below, for the parameter values listed in Table 1.

ρ	0.02	γ	1.5	g	0.02	A_0	2	α	1	β	2	δ	0.00012

Table 1: Simulation parameters for Figure 3

The values of ρ, γ, and g have been chosen as central estimates from the macroeconomics literature.[15] $A_0 = 2$ is chosen so that the value of a statistical life-year at $t = 75$ is 4x consumption per capita, roughly matching Klenow et al. (2024).[16] That is, the first year of the simulation might be taken to denote 1949, when a nuclear war between superpowers first became possible, in which case the 75th year denotes the present. $\bar{\delta}$, α, and β are chosen so that the hazard rate today is approximately 0.1%, matching Stern's (2007) oft-cited figure; and, for clarity in illustration, so that the

[15] We focus on the $\gamma > 1$ case both because it generates the important results and to match evidence from Hall (1988), Lucas (1994), Chetty (2006), and others.

[16] They estimate that this ratio was roughly 5 in the United States in 2019. The figure must be adjusted upward for economic growth since 2019, but downward insofar as we are considering optimal policy across all countries advanced enough to be deploying existentially hazardous technology.

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.

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

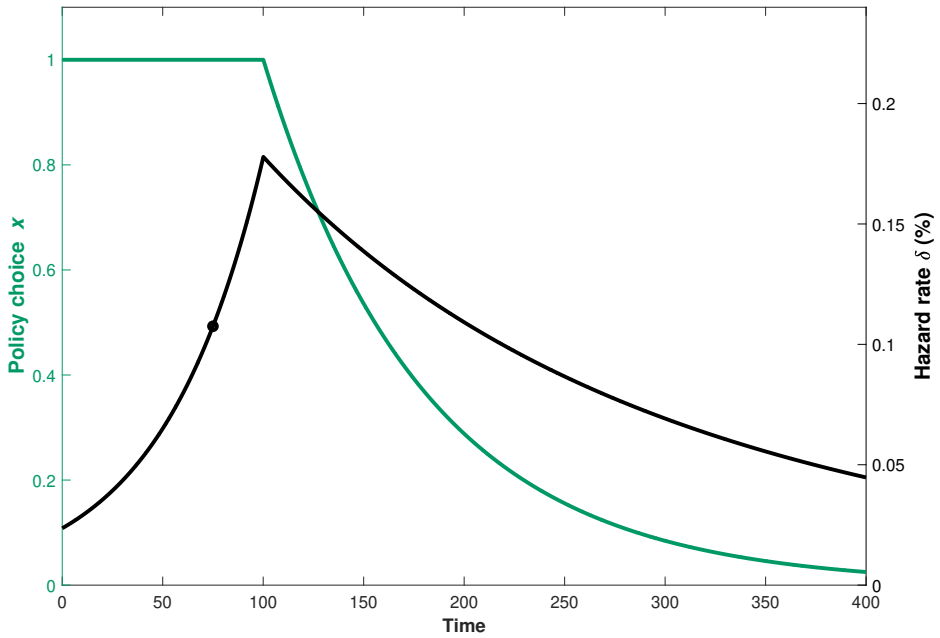


Figure 3: Evolution of policy and hazard along the optimal path

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

As Figure 3 illustrates, one potentially unappealing feature of this simple model is that it implies that, on the optimal path, the hazard rate only rises while no sacrifices whatsoever are made for existential safety. In this it resembles Stokey's (1998) "environmental Kuznets curve", whose damages also rise exponentially with growth and then fall sharply once it becomes optimal to take action.

As in Stokey (1998), this dynamic is 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 at first the hazard rate is allowed to rise. Rising δ can thus be found alongside falling x by tweaking the hazard function around $x = 1$. Such tweaks do not affect the long-run behavior of policy or risk as given by (18)–(20), which are set by the shape of the hazard function around $x = 0$. This is discussed further in Appendix B.4.1.

3.4 Acceleration and state risk

As in the tech-only model of Section 2, the impact on cumulative risk of a temporary shock is ambiguous, but the impact of an acceleration—e.g. a permanent level or growth effect—is always weakly to lower cumulative risk.

Preliminaries — Let $A_{(\cdot)}$ denote the baseline technology path, given by (4). Let $A^* \equiv A_{t^*}$, where t^* is defined as in Proposition 1.

The area under the hazard curve can again be defined by integrating with respect to A instead of t. We will let X denote cumulative risk given that the technology path is $A_{(\cdot)}$ and the policy path x is optimal given $A_{(\cdot)}$:

$$X \equiv \int_0^\infty \bar{\delta} A_t^\alpha x_t^\beta dt = \int_{A_0}^\infty \bar{\delta}\, A^\alpha\, x_A^\beta \left(\frac{dA}{dt}\right)^{-1} dA = \int_{A_0}^\infty \bar{\delta}\, A^\alpha\, x_A^\beta\, \dot{A}_A^{-1}\, dA, \qquad (25)$$

where we will again abuse notation somewhat by letting x_A and $\dot{A}_A$ denote, respectively, the optimal value of x (given technology path $A_{(\cdot)}$) and the value of $\dot{A}$ when the technology level equals the subscripted A.

We will define v_A and δ_A likewise. Note that $\delta_A \equiv \bar{\delta} A^\alpha x_A^\beta$, without dividing this expression by $\dot{A}_A$. That is, δ_A is still a hazard rate: it denotes the probability of catastrophe per unit time at technology level A, not the probability of catastrophe per unit of technological development.

A technology path $\tilde{A}_{(\cdot)}$ that is continuously differentiable almost everywhere and whose right derivative $\dot{\tilde{A}}_{(\cdot)}$ is defined and right-continuous everywhere is an *acceleration* from $\underline{A} \in [A_0, \infty)$ to $\overline{A} \in (A_t, \infty]$ if $\tilde{A}_0 = A_0$ and

$$\dot{\tilde{A}}_A = \dot{A}_A, \;\; A \notin (\underline{A}, \overline{A}); \qquad \dot{\tilde{A}}_A > \dot{A}_A, \;\; A \in (\underline{A}, \overline{A}).$$

The acceleration is *permanent* if $\overline{A} = \infty$ and *temporary* otherwise.

Let $\tilde{A}_{(\cdot)}$ be an acceleration from $\underline{A}$. Define $\tilde{v}_A$ such that at $A < \underline{A}$, $\tilde{v}_A = v_A$, and at $A \geq \underline{A}$, $\tilde{v}_A$ is the costate variable on survival at A given that the subsequent technology path is $\tilde{A}_{(\cdot)}$. Then $\tilde{x}_A$ is defined to equal (13) with $A, \tilde{v}_A$ in place of A_t, v_t; $\tilde{\delta}_A \equiv \delta(A, \tilde{x}_A)$; and $\tilde{X} \equiv \int_{A_0}^\infty \tilde{\delta}_A \dot{\tilde{A}}_A^{-1} dA$.

Given a baseline technology level $\underline{A}$ and a technology growth rate $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$, denote by $\tilde{A}_{(\cdot)}[\epsilon]$ the acceleration from $\underline{A}$ to $\underline{A} + \epsilon$ with

$$\dot{\tilde{A}}_A = \dot{\tilde{A}}, \quad A \in (\underline{A}, \underline{A} + \epsilon).$$

Then the effect on cumulative risk, per unit of technological development, of *instantaneously accelerating to* $\dot{\tilde{A}}$ *at* $\underline{A}$ is defined to be

$$\Delta_{\underline{A}, \dot{\tilde{A}}} \equiv \lim_{\epsilon \to 0} \left(\tilde{X}[\epsilon] - X\right)/\epsilon,$$

where $\tilde{X}[\epsilon]$ is cumulative risk $\tilde{X}$, as defined above, given acceleration $\tilde{A}_{(\cdot)}[\epsilon]$.[17]

Instantaneous level effects — The effect per unit time of a positive shock to the technology level A_t, letting policy adjust instantaneously, depends on whether the shock occurs before or after the regime-change time t^*. At $t < t^*$, temporarily multiplying the technology level by $m > 1$ has no impact on the optimal choice of x.[18] The hazard rate thus rises. The future hazard rate is unaffected, so cumulative risk increases by

$$\delta_t(m^\alpha - 1) > 0$$

per unit of time that the technology level is raised.

At $t \geq t^*$, temporarily multiplying the technology level by $m > 1$ multiplies the policy variable by $m^{-\frac{\alpha+\gamma-1}{\beta+\gamma-1}}$, by (13). In combination, the positive shock to technology and the negative impact on the policy variable multiply the hazard rate by $m^{\alpha-\beta\frac{\alpha+\gamma-1}{\beta+\gamma-1}} = m^{-\frac{(\beta-\alpha)(\gamma-1)}{\beta+\gamma-1}} < 1$. This resulting change in cumulative risk is

$$\delta_t\big(m^{-\frac{(\beta-\alpha)(\gamma-1)}{\beta+\gamma-1}} - 1\big) < 0$$

per unit of time that the technology level is raised.

Instantaneous accelerations — Multiplying the technology growth rate at t by $m > 1$ lowers cumulative risk (per unit of time that the shock lasts) regardless of t. It does so only because the shock decreases the time spent at technology levels around A_t. The shock has no impact on the policy associated with any technology level.

As in the tech-only model, therefore, the impact of this shock on cumulative risk *per unit of increase to the technology level* during the acceleration is

$$\delta_t((m\dot{A}_t)^{-1} - \dot{A}_t^{-1}) < 0.$$

So the impact on cumulative risk *per unit of time* that the acceleration lasts is the above multiplied by the new technology growth rate $m\dot{A}_t$:

$$\delta_t(1 - m) < 0.$$

Accelerations — Consider a "sharp temporary acceleration", in which technology jumps at t from A_t to $\overline{A} > A_t$ and exponential technology growth is then maintained. Since in this model optimal policy is history-independent, this technology shock amounts to a leap forward in time. The resulting change in cumulative risk is

$$-\int_{A_t}^{\overline{A}} \delta_A \dot{A}_A^{-1}\, dA.$$

[17] The effect on an instantaneous acceleration on cumulative risk *per unit time* is $\Delta_{\underline{A},\dot{\tilde{A}}}\dot{\tilde{A}}$, since during the acceleration, $\dot{\tilde{A}}$ units of technology are developed per unit time. This is of the same sign.

[18] Unless m is large enough to reverse inequality (12), a case we will ignore for simplicity.

Such a jump is the limiting case of an acceleration from $\underline{A}$ to $\overline{A}$, which can lower the risk endured at the given range of technology levels for two reasons. First, as in the tech-only model, increasing the technology growth rate at A always lowers cumulative risk directly because the exponent on $\dot{A}_A$ in integral (25) is negative: $\dot{\tilde{A}}_A^{-1} < \dot{A}_A^{-1}$. Second, going beyond the tech-only model, given $A \in [A_t, \overline{A})$, the value of the future at A is higher given faster future technology growth: $\tilde{v}_A > v_A$. By (13), this motivates weakly more stringent policy $\tilde{x}_A \leq x_A$ and thus a weakly lower hazard rate $\tilde{\delta}_A \leq \delta_A$.

Via the first channel alone, the change in cumulative risk achieved by an acceleration is the integral, across technology levels, of the risk reductions achieved by instantaneous accelerations at each technology level:

$$\int_{A_t}^{\overline{A}} \delta_A(\dot{\tilde{A}}_A^{-1} - \dot{A}_A^{-1})dA < 0.$$

Given a policy impact, the cumulative risk reduction achieved is greater.

To summarize:

Proposition 2. ***Acceleration and state risk***
If $\gamma > 1$, so that X is finite, an instantaneous acceleration at $\underline{A}$ to $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$ decreases cumulative risk per unit of technological development during which it endures:

a) $\Delta_{\underline{A},\dot{\tilde{A}}} = \delta_{\underline{A}}(\dot{\tilde{A}}^{-1} - \dot{A}_{\underline{A}}^{-1}) < 0$

and an acceleration $\tilde{A}_{(\cdot)}$ from $\underline{A}$ to $\overline{A}$ decreases cumulative risk by weakly more than the corresponding integral of instantaneous accelerations:

b) $\tilde{X} \leq X + \int_{\underline{A}}^{\overline{A}} \Delta_{A,\dot{\tilde{A}}_A} dA < X$, *with equality strict only if* $\overline{A} \leq A^*$.

The results follow from the integral defining cumulative risk (25) and the definition of instantaneous acceleration. The impacts of shocks to growth on survival are explored in the generalized model of Appendix B.3, and the generalized results are given and proved in detail there in Proposition 7.

Example — Given the parameter values used to illustrate the baseline path are the same as those used to simulate Figure 3, consider a sharp temporary acceleration "today", at $t = 75$, that multiplies A by $e^{0.2} \approx 1.22$, so that at $g = 0.02$, it amounts to a 10-year leap forward.

Recall from the simulation of the previous section 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 then 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 $e^{-\Delta X} \approx 1.011$. In absolute terms S_∞ rises by approximately $0.65 \cdot 0.011 \approx 0.7\%$.

3.5 Discussion

Slow growth makes catastrophe inevitable — As noted in Section 2.2, "deceleration" can render survival impossible. For example, stagnation must do so.

Here, the technology conditions necessary for survival can be stated more precisely. Suppose $\gamma > 1$ and consider a permanent deceleration after which technology grows power-functionally, so that $\tilde{A}_t = t^k$ for some $k > 0$. The exponential growth rate of $\tilde{A}$, denoted $\tilde{g}$, is then time-varying, with $\tilde{g}_t = k/t$. By (15) and since $\tilde{g}_v \to 0$, δ_t then falls to 0 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 iff

$$k > \frac{\beta + \gamma - 1}{(\alpha - \beta)(\gamma - 1)}. \tag{26}$$

Growth vs. patience — Faster growth increases the willingness to pay for safety. Those concerned about the safety of the long-term future often pursue the same result via ethical arguments for a low rate of time preference. Consider e.g. the Stern–Nordhaus debate (and the long debate since) over the discount rate to use in climate policy, or the arguments for concern for the future made by philosophers such as Parfit (1984), Cowen and Parfit (1992), Ord (2020), and MacAskill (2022).

Mere level effects—permanent proportional increases to A—here impact policy and survival probability similarly to proportional decreases to ρ. In brief, this is because proportionally decreasing ρ raises v by a similar proportion (c.f. (17)), and by (13), proportional increases to v and to A have similar effects on policy.

Acceleration can lower life expectancy — If $\gamma < 1$, so that civilization's "life expectancy" is finite, accelerations can decrease it. To see this, recall that stagnation at low A yields a permanent hazard rate of $\bar{\delta}A^\alpha$. This may be arbitrarily low, so the expected duration until catastrophe $1/(\bar{\delta}A^\alpha)$ may be arbitrarily high. When $\gamma < 1$, an acceleration can quickly yield a hazard rate that permanently approximates δ^* (24). Civilizational life expectancy can thus fall to approximately $1/\delta^*$.

4 Transition risk

4.1 Motivation

A hazard function of the form $\delta(A_t, x_t)$ captures what we have 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 (2024) 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 is posed by discovering and deploying new technologies with unknown consequences. If so, given a positive-growth baseline, does accelerating technological development further increase cumulative risk?

4.2 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. We will again restrict our consideration to a constant elasticity hazard function:

$$\delta_t = \bar{\delta} A_t^{\alpha} \dot{A}_t^{\zeta} x_t^{\beta}, \quad \bar{\delta} > 0,\ \zeta \geq 0,\ \beta > 1. \tag{27}$$

Hazard function (5) is the special case of (27) with $\zeta = 0$ (and $\beta > \alpha > 0$). If $\zeta > 0$, however, the model is most naturally interpreted as one in which risk is posed by the introduction of new technologies—"draws from Bostrom's urn"—which each increase A by one unit. Fixing policy, introducing multiple technologies can pose more, less, or equal risk if done concurrently than if done in sequence, depending on the sign of $\zeta - 1$. Introducing more advanced technologies can pose more, less, or equal risk than introducing less advanced technologies, depending on the sign of α.

Alternatively, to interpret one "new technology" as a *proportional* increase to A, simply rewrite the hazard function as

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

Here $\alpha + \zeta > 0$ is the condition under which developing more advanced technologies poses more risk than developing less advanced technologies. If $\dot{A}/A$ has long been roughly constant, the view that the hazard rate has risen must be attributed to the increasing danger of each "technological development" in this sense.

Finally, consider the case of $\alpha = -1$, $\zeta = 1$, so that

$$\delta_t = \bar{\delta} \frac{\dot{A}_t}{A_t} x_t^{\beta}.$$

Here, fixing x, each proportional increase to A induces the same hazard, independently of how quickly the increase occurs. In the absence of policy—with $x = 1$ (or any other constant) permanently—this model is essentially equivalent to the "Russian roulette" model of Jones (2016)[19] and the AI risk model of Jones (2024).

4.3 Acceleration and transition risk

Without policy — Suppose that the baseline technology path $A_{(\cdot)}$ is continuously differentiable, with a positive derivative. Let $\hat{A} \equiv \lim_{t\to\infty} A_t$ be finite or infinite.

As implied above, fixing policy, whether acceleration increases or decreases cumulative risk depends on whether ζ is greater or less than 1. This can, again, be seen by integrating the hazard curve with respect to A:

$$X = \int_0^\infty \bar{\delta} A_t^\alpha \dot{A}_t^\zeta dt = \int_{A_0}^{\hat{A}} \bar{\delta} A^\alpha \dot{A}_A^{\zeta-1} dA.$$

Given acceleration $\tilde{A}_{(\cdot)}$ from $\underline{A} \in [A_0, \hat{A})$ to $\overline{A} \in (\underline{A}, \hat{A}]$, cumulative risk equals

$$\tilde{X} = X + \int_{\underline{A}}^{\overline{A}} \bar{\delta} A^\alpha \big(\dot{\tilde{A}}^{\zeta-1} - \dot{A}_A^{\zeta-1}\big) dA.$$

The integral is negative if $\zeta < 1$, zero if $\zeta = 1$, and positive if $\zeta > 1$.

Because the Jones models implicitly adopt $\zeta = 1$, though there is a technology level $\hat{A} < \infty$ at which it is optimal to halt technological development (Appendix B.5), the speed of technological development before $\hat{A}$ does not affect cumulative risk.

With policy — We have seen that the impact of acceleration on cumulative transition risk is ambiguous absent policy. Under optimal policy it remains ambiguous, but a tendency for acceleration to lower cumulative risk may be reintroduced.

For simplicity and focus, we will assume that A grows at a constant rate g and that $\gamma > 1$. Also, since given exponential growth $g_{\dot{A}} = g$, we will impose

$$\beta > \alpha + \zeta, \tag{28}$$

which, rather than $\beta > \alpha$, is now the condition necessary for survival without $C_t = A_t x_t \to 0$. Under these conditions, since $\dot{A}$ is proportional to A, the planner's problem is precisely as described in Section 3.3, with $\alpha + \zeta$ taking the place of α (up to a coefficient g^ζ that can be incorporated into $\bar{\delta}$). Baseline x and δ paths, and S_∞, are unchanged. The existential risk Kuznets curve remains.

Let A^* denote the uppermost technology level at which it is optimal to set $x = 1$ on the baseline technology path. Since the first-order condition

$$\frac{\partial u}{\partial x_t}(A_t, x_t) \geq \frac{\partial \delta}{\partial x_t}(A_t, x_t) v_t \quad \Longrightarrow \quad A_t^{1-\gamma} x_t^{-\gamma} \geq \bar{\delta} A_t^\alpha \beta x_t^{\beta-1} v_t$$

[19] Our $\bar{\delta}$ is the variable there denoted π.

must be satisfied everywhere and hold with equality for $x < 1$, we have

$$x_A = \begin{cases} 1 & A \leq A^*, \\ \left(\bar{\delta}\beta A^{\alpha+\gamma-1}\dot{A}_A^{\zeta} v_A\right)^{-\frac{1}{\beta+\gamma-1}} & A > A^*. \end{cases} \tag{29}$$

Substituting (29) into the expression for cumulative risk

$$X = \int_{A_0}^{\infty} \bar{\delta}\, A^{\alpha} \dot{A}_A^{\zeta-1}\, x_A^{\beta}\, dA, \tag{30}$$

we have

$$X = \int_{A_0}^{A^*} \bar{\delta} A^{\alpha} \dot{A}_A^{\zeta-1} dA + \int_{A^*}^{\infty} \left(\bar{\delta}^{1-\gamma}\beta^{\beta} A^{(\beta-\alpha)(\gamma-1)} v_A^{\beta}\right)^{-\frac{1}{\beta+\gamma-1}} \dot{A}_A^{\zeta\frac{\gamma-1}{\beta+\gamma-1}-1} dA. \tag{31}$$

Recall that a technology path $\tilde{A}_{(\cdot)}$ is an acceleration if $\dot{\tilde{A}}_A > \dot{A}$ for technology levels $\underline{A} \in [A_0, \infty)$ to $\overline{A} \in (\underline{A}, \infty]$. With or without policy, an acceleration affects cumulative risk directly, by changing the technology growth rate from $\underline{A}$ to $\overline{A}$. With policy, an acceleration also affects cumulative risk indirectly by affecting v_A for $A \in [\underline{A}, \overline{A})$, which affects policy at this range of technology levels.

Under hazard function (5), faster technology growth is always preferred, as explained following (12). A future with faster growth is more valuable, so an acceleration from $\underline{A}$ to $\overline{A}$ raises v_A for $A \in [\underline{A}, \overline{A})$. Under hazard function (27), this no longer holds. Unlike an increase to A_t, an increase to $\dot{A}_t$ brings no contemporaneous benefit, but it now imposes risks that can still be mitigated only with less contemporaneous consumption. We can see that growth is sometimes undesirable most plainly when $\alpha = -1$, $\zeta = 1$: again, this is the Russian roulette model, and as Jones finds, with $\gamma > 1$, the planner grows technology only to a finite level.[20]

These complexities are avoided when we focus on instantaneous accelerations. The impact of an acceleration from $\underline{A}$ to $\overline{A}$ on v_A, for $A \in [\underline{A}, \overline{A})$, falls to zero as $\overline{A} - \underline{A} \to 0$. The impact of a *brief* acceleration on cumulative risk is therefore approximately the impact found when we ignore impacts on v_A.

Proposition 3. ***Instantaneous acceleration and transition risk***
Given hazard function (27) and technology path (4), choose a technology level $\underline{A} > 1$ and growth rate $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$. If

a. *$\underline{A} \geq A^*$ and $\zeta < (=, >)\, 1 + \frac{\beta}{\gamma-1}$, or*

b. *$\underline{A} < A^*$, $\zeta < (=, >)\, 1$, and $\dot{\tilde{A}}$ maintains (29) $= 1$ at $A = \underline{A}$,*

[20] In this more general model, the optimality of *stagnation* is knife-edge (Appendix B.5), but the result that accelerating from $\underline{A}$ to $\overline{A}$ does not necessarily raise v_A for $A \in [\underline{A}, \overline{A})$ is not.

then $\Delta_{\underline{A},\dot{\bar{A}}} < (=, >)\, 0$.

The result follows essentially immediately from the exponent on $\dot{A}_A$ in (30).[21] In particular, instantaneous acceleration after A^* lowers cumulative risk as long as

$$\zeta \frac{\gamma - 1}{\beta + \gamma - 1} - 1 < 0 \implies \zeta < 1 + \frac{\beta}{\gamma - 1}. \tag{32}$$

It is sufficient, though not necessary, for (32) 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, by (28), $\zeta < \beta + 1$, so $\frac{\zeta}{\beta+1} < 1$. Since macroeconomic estimates of $\gamma \leq 2$ are standard, this result suggests that accelerations lower cumulative risk on the optimal path in the context of transition risk, at least if they occur late enough in time that mitigation is already underway.

Again, this is without considering the fact that an increase to future growth can change the value of the future. Though the direction of this change is ambiguous, it is often taken for granted that, on a conventional discount rate, faster technology growth would currently be a benefit. If so, this is another channel through which a (positive-duration) acceleration increases concern for safety.

It may be counterintuitive that instantaneous acceleration reduces risk only when γ lies *below* a bound. The result is due to the fact that, when γ is high, the marginal utility of consumption rises rapidly as x is cut, so following an acceleration, a small cut to x suffices to equalize the marginal utility of consumption with that of safety spending. The higher γ is, the more quickly x falls as A rises, but the *less* sensitive x is to a change in $\partial\delta/\partial x$—e.g. an increase due to higher $\dot{A}$—at a given value of A.

4.4 Discussion

Nonrivalry in safety effort — Hazard function (27) is explored here mainly for its simplicity and similarity to (5). This functional form overemphasizes a channel through which the risks posed by a series of technological developments can be cheaper to mitigate if they occur at once than if they occur in sequence. Suppose that $\beta \approx 1$, that $\zeta = 1$, and that two small A-increases—"experiments"—can occur in sequence or simultaneously. If they occur in sequence, halving the risk posed by each requires halving x and thus consumption for two periods in a row. If they occur simultaneously, the same risk-reduction only requires halving consumption for one period.

For some kinds of experiments and safety efforts, this "nonrivalry" assumption is reasonable. Monitoring wastewater to detect pandemics early reduces the risk posed by the relevant biological experiments by a proportion independent of how many are

[21] A rigorous proof may be found in Appendix C.1.

underway. For other kinds, the assumption is not reasonable: e.g. it does not apply to the safety equipment that must be used at each lab (c.f. Appendix B.4.2).

This model is thus in no sense a thorough study of the relationship between growth and transition risk. It is intended only to offer two limited lessons. First, under optimal policy, the effect of acceleration on transition risk remains ambiguous. Second, the presence of an optimal policy response can change the conditions under which acceleration lowers risk, and can relax them to the extent that safety efforts are nonrival across contemporaneous risks.

Stagnation vs. deceleration — When $\zeta > 0$, complete stagnation ($\dot{A} = 0$) is always the safest path of all. Nevertheless, we have seen with and without policy that given a positive growth rate, faster growth can decrease risk.

This is because, given stagnation at $\hat{A}$, levels $A > \hat{A}$ are never attained. Cumulative risk is thus not (30) but (30) with the ∞ replaced by $\hat{A}$. Absent stagnation, all levels of A are attained; the growth rate only determines the risk endured at each. The direct cost of faster progress during a given A-range (higher risk per unit time) 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 familiar fact that when progress is faster we do not linger in a given A-range as long.

5 Conclusion

Human activity can create or mitigate existential risks. The framework presented here illustrates that, under various conditions, existential risk should be expected to exhibit a Kuznets curve. This observation offers a potential economic explanation for the claim that we are in a "time of perils". We may be advanced enough to be able to destroy ourselves, but not yet rich 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 have massive long-term consequences.

This framework also highlights a channel through which some efforts intended to reduce existential risk may backfire. In the absence of policy, when risk is posed by the *existence* of advanced technologies, broad-based decelerations to technological development typically worsen or do not affect the odds of long-term survival. Given an optimal policy response, even by a policymaker with little concern for the long-term future, this impact is magnified. The impact 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 stagnation can make a catastrophe inevitable that might otherwise have been avoided.

This lesson comes with three caveats. First, it is not an argument against regulating the use of risky technologies. Indeed, a primary channel explored here through which technological development lowers risk is that it hastens the day when regula-

tion is severe. Some recent reactions to calls for heavy AI regulation, 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 reasons presented in this paper.

Second, when risk is posed by the *development* of advanced technologies, the effect of acceleration on risk is ambiguous. In the "transition risk" models of Jones (2016, 2024), acceleration does not affect cumulative risk. Under slight modifications to these models, the impact may be positive or negative. Policy may facilitate a tendency for acceleration to weakly decrease cumulative risk, as illustrated in Section 4; but it seems likely that in other plausible models it would not.

Third, where we have found that policy magnifies a negative link between acceleration and cumulative risk, we have assumed that policy is optimal. If it is not, then the impact of acceleration on cumulative risk may be reduced or even overturned, as illustrated in Section 3.1.[22] The appropriate lesson about the impact of policy on the relationship between acceleration and risk is only that, to the extent that the policy regime equates or will eventually equate the marginal utility of consumption to the marginal expected discounted utility of safety expenditure, consumption-increasing technological development has the unseen benefit of speeding future safety efforts. For slowing technological development to lower cumulative risk, there must be a policy failure severe and lasting enough to outweigh this potentially large benefit.

In this light, further research on the nature of policy distortions around the regulation of risky technologies would be valuable. Exploring the long-term implications of other models of anthropogenic existential risk, and of optimal policy in the face of it, could be valuable as well, to better characterize the scope of the result that optimally regulated acceleration weakly lowers cumulative risk. If plausible models are found under which the result is overturned, this will naturally pose important questions 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 broad-based technological progress are sufficiently small.

[22]Shulman and Thornley (2024) argue that the policy response to existential risk is in fact far from optimal, even under a conventional discount rate.

A Appendix

A.1 Existence and uniqueness of optimal policy path

Necessary and sufficient conditions — The optimization problems analyzed in Sections 3–4, and the supplemental appendices, feature one choice variable x and one state S. Expected flow utility at t is $S_t u(A_t, x_t)$ for a $\mathcal{C}^2$ function $u(\cdot)$, strictly concave in x, with a lower Inada condition on x. The law of motion for S is $-S_t\delta(A_t, \dot{A}_t, x_t)$ for a $\mathcal{C}^2$ function $\delta(\cdot)$. A, $\dot{A}$ are independent of x, so operate as functions of t.

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

$$\mathcal{L}(S_t, x_t, v_t, \mu_t, t) = S_t u(x_t, t) - v_t S_t \delta(x_t, t) + \mu_t(1 - x_t) \tag{33}$$

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

(33) satisfies the Mangasarian concavity condition that $\mathcal{L}(\cdot)$ is weakly concave in S and x. So applying Caputo (2005), Theorems 14.3-4 and Lemma 14.1,[23] 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 semi-continuity of x and S, only if—for some semi-differentiable path of v and some semi-continuous path of $\mu \geq 0$, at all t the following first-order conditions are satisfied

$$\frac{\partial \mathcal{L}}{\partial x_t}(S_t, x_t, v_t, \mu_t, t) = \mu_t \frac{\partial \mathcal{L}}{\partial \mu_t}(S_t, x_t, v_t, \mu_t, t) = 0, \quad \frac{\partial \mathcal{L}}{\partial \mu_t}(S_t, x_t, v_t, \mu_t, t) \geq 0, \tag{34}$$

as well as the transversality condition that

$$\lim_{t\to\infty} e^{-\rho t} v_t = \lim_{t\to\infty} e^{-\rho t} v_t S_t = 0. \tag{35}$$

Given optimal paths of x and S and corresponding paths of v and μ, v is continuous and satisfies

$$\dot{v}_t = \rho v_t - \frac{\partial \mathcal{L}}{\partial S_t} = \rho v_t - u(x_t, t) - v_t \dot{S}_t = (\rho + \delta(x_t, t))v_t - u(x_t, t) \tag{36}$$

except at discontinuity points of x, where v's right and left derivatives may differ.

The transversality condition — Given a continuous v path, only

$$x_t = \begin{cases} 1, & \frac{\partial u}{\partial x}(1, t) - \frac{\partial \delta}{\partial x}(1, t)v_t \geq 0; \\ x_t \,:\, \frac{\partial u}{\partial x}(x_t, t) - \frac{\partial \delta}{\partial x}(x_t, t)v_t = 0, & \text{otherwise} \end{cases} \tag{37}$$

[23] 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.

$$\mu_t = \frac{\partial u}{\partial x_t}(x_t, t) - \frac{\partial \delta}{\partial x_t}(x_t, t) v_t \tag{38}$$

satisfy (34) for all t. Any such x path is well-defined, by the continuous differentiability of $u(\cdot)$ and $\delta(\cdot)$ in x and the fact that $u(\cdot)$ and $\delta(\cdot)$ strictly increase in x. Any such x path is also right-continuous in time, by

- the twice continuous differentiability of $u(\cdot)$ and $\delta(\cdot)$ (expressed as functions of x, A, and for $\delta(\cdot)$ in Section 4, $\dot{A}$);
- the right-continuity of the right derivative of $A_{(\cdot)}$ in time;
- (for Section 4, given exponential growth or instantaneous acceleration) the right-continuity of the right derivative of $\dot{A}_{(\cdot)}$;

and the implicit function theorem. Any such μ path is then also right-continuous in time by the composition of continuous functions. To show there exists an optimal path, and that only one such path is semi-continuous, it will now suffice to show that there is a unique v path for which (35)–(36) are satisfied given the corresponding x path (37) and its implied S path, and that the corresponding x path is semi-continuous (in fact it is right-continuous).

The solution to differential equation (36) is

$$v_t = e^{\int_0^t (\rho + \delta_s) ds} \left(v_0 - \int_0^t e^{-\int_0^s (\rho + \delta_q) dq} u(x_s, s) ds \right) \tag{39}$$

$$\implies v_0 = \int_0^t e^{-\rho s} S_s u(x_s, s) ds + e^{-\rho t} S_t v_t. \tag{40}$$

Since (40) is continuous in t (by the boundedness of $u(\cdot)$ and the continuous evolution of S) and holds for all t, v satisfies (35)–(36) iff

$$v_0 = \int_0^\infty e^{-\rho t} S_t u(x_t, t) dt. \tag{41}$$

Given (37), v_t determines x_t for all t. Given (36), v_t and x_t determine the right derivative of v for all t. Given v_0, therefore, there is a unique path of v—and thus of x, and thus of S—compatible with (36)–(37). We will now show that there is at least one value of v_0 for which (41) is satisfied, given the corresponding x and S paths. For such a v_0, the corresponding variable paths by construction satisfy (34)–(35).

Existence — Let $v(v_0)$ and $x(v_0)$ denote the unique paths of v and x compatible with (36)–(37) for which $v_0(v_0) = v_0$. By (39), $\lim_{v_0 \to -\infty} v_t(v_0) = -\infty$ for all $t \geq 0$. By (37), 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 (36), 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 integrating

(36), with (37) substituted for x_t, would allow us to identify $v_0 = \tilde{v}_0$. 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 (41) 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 (41) are upper-bounded. So, for $\overline{v}_0 > \overline{U}$, the left-hand side of (41) exceeds the right-hand side.

By (37) and the implicit function theorem, x_t is continuous in v_t for all t. (36) then implies that $\dot{v}_t$ is defined and continuous in v_t for all t, and thus that $v_t(v_0)$, then $x_t(v_0)$, and then the right-hand side of (41) are 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 (41) holds.

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

Uniqueness (among semi-continuous x paths) also follows from the facts that a path is optimal iff v_0 attains its maximum feasible value and that, given (34)–(35), v_0 determines a unique path for every variable.

A.2 Proof of Proposition 1

Asymptotic constancy of g_v — From (36), because v is the costate on S, it obeys

$$\dot{v}_t = (\rho + \delta_t)v_t - u(C_t) \quad \Longrightarrow \quad g_{vt} = \rho + \delta(A_t, x_t) - \frac{u(A_t x_t)}{v_t}. \tag{42}$$

Let $\tilde{\beta} \equiv \beta + \gamma - 1$. From (13), once x_t is interior we have

$$x_t = A_t^{-\frac{\alpha+\gamma-1}{\tilde{\beta}}} \left(\bar{\delta}\beta v_t\right)^{-\frac{1}{\tilde{\beta}}}. \tag{43}$$

Substituting (43) into (42) yields

$$g_{vt} = g_v(v_t, t) \equiv \begin{cases} \rho + KA_t^{\frac{(\beta-\alpha)(1-\gamma)}{\tilde{\beta}}} v_t^{-\frac{\beta}{\tilde{\beta}}} + \frac{1}{1-\gamma} v_t^{-1}, & \gamma \neq 1; \\ \rho + \log\left(A_t^{-\frac{\beta-\alpha}{\beta}} \left(\bar{\delta}\beta v_t\right)^{-\frac{1}{\beta}}\right) v_t^{-1}, & \gamma = 1, \end{cases} \tag{44}$$

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 (44) 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

$$\lim_{t\to\infty} g_{vt} = \rho + \frac{1}{v^*(1-\gamma)}. \tag{45}$$

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 (44) 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 (7), ρ is higher, and is ruled out as a steady state by the transversality condition (35). Then because the limit

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

is defined and continuous in g_v, we must have

$$\lim_{t\to\infty} g_{vt} = \frac{(\beta-\alpha)(1-\gamma)}{\beta} g. \tag{46}$$

Otherwise $g_v \to -\infty$, 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 (44) with respect to time yields $\dot{g_{vt}}$ strictly, 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 $t \to \infty$. So

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

are defined and continuous in g_v, and $g_v \not\to \infty, -\infty$ imply $\lim_{t\to\infty} g_{vt} = 0$.

Asymptotic behavior of other variables — With the asymptotic behavior of g_v pinned down, that of x and C follows immediately, as does that of δ if $\gamma > 1$.

To find the asymptotic behavior of δ given $\gamma \leq 1$, rearrange (44) to get

$$v_t = \frac{u(C_t)}{\rho + \delta_t - g_{vt}}, \tag{47}$$

and substitute (47) into $C_t^{1-\gamma} = \delta_t \beta v_t$ ((11) rearranged) to get

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

If $\gamma < 1$, the limit of g_v (46) and $C \to \infty$ from (22) imply

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

If $\gamma = 1$, substitute 0 for g_{vt}. By (22), $\exists \underline{C} > 0 \ : \ \lim_{t\to\infty} \frac{C_t}{e^{\frac{\beta-\alpha}{\beta}gt}} = \underline{C}$, so

$$\lim_{t\to\infty} \delta_t t = \lim_{t\to\infty} \frac{\rho - g_{vt}}{\beta\big(\log(C_t/e^{\frac{\beta-\alpha}{\beta}gt}) + \log(e^{\frac{\beta-\alpha}{\beta}gt})\big)/t - 1/t} = \frac{\rho}{(\beta-\alpha)g}.$$

References

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

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

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

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

_ , *Superintelligence: Paths, Dangers, Strategies*, Oxford University Press, 2014.

_ , "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).

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.

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," 2024. Working paper.

Klenow, Peter J., Charles I. Jones, Mark Bils, and Mohamad Adhami, "Population and Welfare: The Greatest Good for the Greatest Number," June 2024.

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.

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.

_ , "Robust longterm comparisons," 2024. Blog post.

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.

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

Online appendix

B Supplemental materials

B.1 Calibrating the elasticity of the hazard rate to safety expenditures

Shulman and Thornley (2024) estimate that well-targeted expenditures of $400B over the next decade would reduce the probability of existential catastrophe over the next decade by at least 0.1% in absolute terms, from a baseline of 1.85%.

The scale of the magnitude of the risk is taken from Ord's (2020, p. 167) educated guesses and may be disputed. However, an estimate of β depends only on the *proportion* by which a given consumption sacrifice will reduce the hazard rate. We will rely on Shulman and Thornley's assessment that expenditures of $400B would multiply the probability of existential catastrophe over the next decade by at most

$$1 - \frac{0.1\%}{1.85\%} \approx 0.946, \tag{48}$$

while remaining agnostic about the the magnitude of the probability. For instance, we are trusting their assessments of the extent to which disease monitoring expenditures would be able to prevent existentially hazardous anthropogenic pandemics by helping authorities to contain them early, while remaining agnostic about the probability per year that such a pandemic will arise.

Global consumption per year is currently approximately $72.5T.[24] If real consumption grows at 2% per year and the relevant interest rate is 5% per year, the present value of global consumption over the next ten years is approximately $\$72.5\text{T} \times (1 - e^{-10(0.05-0.02)})/(0.05 - 0.02) \approx \$626.4\text{T}$. A sacrifice of $\$400\text{B} = \$0.4\text{T}$ in today's dollars over the next decade is thus a sacrifice that multiplies consumption by a fraction of

$$1 - \frac{0.4}{626.4} \approx 0.99936. \tag{49}$$

Given $x^\beta < 0.946$ at $x \approx 0.99936$, it follows that

$$\beta > \frac{\log(0.946)}{\log(0.99936)} \approx 86.7.$$

This exercise of course tells us nothing about whether it is reasonable to assume a constant-elasticity hazard function in general. If the Shulman and Thornley estimate

[24] World Bank national accounts data and OECD National Accounts data files: Final consumption expenditure (current US$). Retrieved from `https://data.worldbank.org/indicator/NE.CON.TOTL.CD?end=2022&start=2022`, May 20, 2024.

is correct within three orders of magnitude, however, it does prove that the hazard function is currently convex over at least some range of feasible consumption levels. This follows immediately from the facts that (49) > (48) and that the hazard rate cannot be cut by a proportion greater than one.

B.2 State risk with policy: growth vs. patience

In the context of the model of Section 3, we will compare the effects on cumulative risk of a sharp and permanent level effect at t, in which A is multiplied by m slightly greater than 1, with the effects of permanently dividing ρ by m. If $\gamma \leq 1$, the similarity of the two interventions is trivial: cumulative risk is infinite both before and after each intervention. We will thus assume $\gamma > 1$.

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

Before t^*, therefore, the impact of a level effect on cumulative risk rises exponentially with δ_t. Early in time δ_t may be arbitrarily low, so the impact of the level effect on cumulative risk may as well. 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 $v_t \to \bar{v}$. By (13), in the limit,

$$x_t \approx \left(\bar{\delta}\beta\bar{v}\right)^{-\frac{1}{\beta+\gamma-1}} A_t^{-\frac{\alpha+\gamma-1}{\beta+\gamma-1}}. \tag{50}$$

At large t, permanently multiplying A by $m > 1$ multiplies x_s, at each $s \geq t$, 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

$$\begin{aligned}(\beta-\alpha)(\gamma-1) &= \beta \\ \iff \gamma &= 2 + \frac{\alpha}{\beta-\alpha},\end{aligned} \tag{51}$$

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 (51) 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.

B.3 State risk with policy: generalized results

Sections 3.3–3.4 are set in the environment of Section 3.2. The three components of this environment are the technology path, the function from technology and policy to the hazard rate, and the utility function. A functional form is assumed for each.

Here we will maintain CRRA utility with $\gamma > 1$. We will however greatly relax our assumptions on the technology path and the hazard rate, to identify the conditions under which the lessons of Sections 3.3–3.4 are maintained.

In Sections B.3.2–B.3.3, generalizing Proposition 1 from Section 3.3, 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. A central result is that, unless 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 B.3.4, generalizing Section 3.4, we likewise find that when a hazard function is compatible with survival, faster technology growth generally increases the probability of survival. The results support the robustness of the lessons drawn from hazard function (5): that survival is likely possible on the optimal path, and that faster consumption technology growth, if optimally regulated, will raise its probability.

B.3.1 Assumptions

Assumptions on technology growth — We will assume throughout only that the technology path $A_{(\cdot)}$ satisfies some or all of the following conditions:

A1. continuous differentiability almost everywhere and right-continuity of the right derivative $\dot{A}$ everywhere, with $\dot{A}_t > 0$ for all t;
A2. $A_0 > 1$;
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.

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's policy.

Return to the three desiderata preceding the introduction of hazard function (5). We will assume weakenings of two of these desiderata directly, and certain results will

require a weakening of the third. In particular, we will 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 $\delta > 0$.[25]

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,[26]
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.

Note that the constant elasticity hazard function of Sections 3.2–3.4 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: we allow new technologies to lower the hazard rate at a given degree of foregone consumption.

B.3.2 The end of consumption growth

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

Given hazard function (5), $C^* = \infty$, by (19) from Proposition 1. However, some admissible hazard functions motivate decreases to x fast enough that we do not have $C^* = \infty$. C^* may be finite, or C_t may oscillate indefinitely.

Proposition 4. ***The end of consumption growth***
Given an admissible hazard function $\delta(\cdot)$, define

$$R(C) \equiv \lim_{A\to\infty} \frac{\partial\delta}{\partial x}\Big(A, \frac{C}{A}\Big) \frac{C^\gamma}{A}\bar{v}, \tag{52}$$
$$R^* \equiv \lim_{C\to\infty} R(C).$$

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 C.2. □

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

[26] 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.

To interpret the result, recall that $x = C/A$. (52) 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.

Recall from (17) that $\bar{v} \equiv \frac{1}{\rho(\gamma-1)}$. When $R(C) > 0$, therefore, $R(C)$ decreases in ρ. A lower discount rate 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. Consumption stagnation is not in general desirable when ρ is sufficiently low, or undesirable when ρ is sufficiently large: for many hazard functions, as shown at the end of the next subsection, R^* is above 1 (even infinite) or below 1 (even 0) for any $\rho > 0$. Still, Proposition 4 illustrates how calls for an "end to growth" may be compatible with this model. Concern for the future can motivate controls on technological deployment strict enough to halt growth in *consumption*, despite the tendency for accelerating *technological development* to lower cumulative risk.

B.3.3 The Kuznets curve generalized

Proposition 5. ***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 C.3. 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:

$$A_t^{1-\gamma} x_t^{-\gamma} \geq \frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t$$
$$\implies (A_t x_t)^{1-\gamma} \geq \frac{\partial \delta}{\partial x}(A_t, x_t) x_t\, v_t. \tag{53}$$

If $C^* = \infty$, the left-hand side of (53) 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 (53) 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 4.

Proposition 6. ***Survival generalized***
Given an admissible hazard function $\delta(\cdot)$ and an admissible technology path $A_{(\cdot)}$ such that, for some $k > 1$ and some $\underline{t}$ we have

$$A_t \geq t^{\frac{k}{\gamma-1}} \quad \forall t > \underline{t}, \tag{54}$$

define

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

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 C.4. □

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

$$C_t \propto t^{\frac{k}{\gamma-1}}. \tag{55}$$

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 (55) 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 (55) for $k = 1$.

If C_t grows more quickly than (55) for some $k > 1$, then the left-hand side of (53) 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 (55) for $k = 1$, then the left-hand side of (53) 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 3.2 for the case of exponential growth at rate g.

$$\begin{aligned}\tilde{R}(k) &= \lim_{t\to\infty} \bar{\delta} e^{\alpha g t} \beta \Big(\frac{t^{\frac{k}{\gamma-1}}}{e^{gt}} \Big)^{\beta-1} \frac{t^{\frac{k\gamma}{\gamma-1}}}{e^{gt}} \bar{v} \\ &= \bar{\delta}\beta\bar{v} \lim_{t\to\infty} e^{-(\beta-\alpha)gt} t^{\frac{\beta+\gamma-1}{\gamma-1}k} = 0 \end{aligned} \tag{56}$$

for any k, since $\beta > \alpha$. So $\lim_{k\downarrow 1} \tilde{R}(k) = 0 < 1$. Part (a) of Proposition 6 thus generalizes the conclusion of (26) that, with hazard function (5), 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, (56) $= \infty$ 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 given exponential technology growth (and indeed given any $A_{(\cdot)}$ that is eventually bounded above zero).

B.3.4 Acceleration and state risk generalized

For any admissible hazard function, the lessons of Section 3.4 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.

Proposition 7. ***Acceleration and state risk generalized***
Choose an admissible technology path $A_{(\cdot)}$ and hazard function $\delta(\cdot)$.
Given $\underline{A}$, $\dot{\tilde{A}}$ with $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$,

a) $\Delta_{\underline{A},\dot{\tilde{A}}} = \delta_{\underline{A}}(\dot{\tilde{A}}^{-1} - \dot{A}_{\underline{A}}^{-1}) < 0$.

Given an acceleration $\tilde{A}_{(\cdot)}$ from $\underline{A}$ to $\overline{A}$,

b) *If $X < \infty$, then $\tilde{X} \leq X + \int_{\underline{A}}^{\overline{A}} \Delta_{A,\dot{\tilde{A}}_A} dA < X$.*

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

Proof. See Appendix C.5. □

The intuition is the same as illustrated in Section 3.4. Acceleration in effect horizontally rescales all or part of the hazard curve by leaving less time spent at each state. It may also induce more stringent policy at each state, in which case the weak inequality of part (b) is strict.

B.3.5 Discussion

Accelerations vs. level effects — Given a technology path $A_{(\cdot)}$ satisfying A1 and A4, say that a differentiable technology path $\tilde{A}_{(\cdot)}$ is a level effect to $A_{(\cdot)}$ (at time 0) if

$$\exists m > 1 \ : \ \tilde{A}_t = mA_t \ \ \forall t.$$

When technology growth is exponential, level effects are (sharp) temporary accelerations. Otherwise, they may be distinct.

Unlike temporary accelerations, level effects do not always decrease cumulative risk outside the exponential growth context. Consider for example hazard function (5) with a technology path $A_{(\cdot)}$ that is nearly stagnant for an arbitrarily long period, say for $t \leq 99$; that grows exponentially at $t > 99$; and for which the implied regime-change time is $t^* = 100$. A level effect—a jump in the technology level at $t = 0$—then raises the technology level during the arbitrarily 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 $t = 99$.

The direction of technical change — 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 and has some effect on the hazard rate at any given level of consumption. In practice, however, technological development is surely at least somewhat *directed*: tradeoffs between consumption and risk in later periods are affected by the extent to which policymakers and firms in earlier periods have developed 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.

In positing a baseline sequence of maximum potential consumption levels $\{A_t\}$ and a hazard function $\delta(A, \cdot)$, we are simply describing a path of possibilities frontiers over time, not embedding any assumptions about how this path is generated. In particular, we are not assuming that there is only one way it is possible for technology to unfold. If we posit a wider space of possible production technologies than the sequence adopted on the baseline path, we must simply clarify that our results only pertain to "accelerations" in the sense of increases to the rate of motion along the baseline path. 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.[27]

In the next section (Appendix B.4), the lessons of this generalized model are used to explore two particular hazard functions 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.

B.4 State risk with policy: Two more hazard functions of interest

We will assume that technology grows at a constant rate $g > 0$.

B.4.1 A lower Inada condition on safety

As shown in Section 3.3, 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. This result is arguably at odds with the experience of the last century, during which the hazard rate has arguably risen while existential safety expenditures have risen (from essentially 0). 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): 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". 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 optimal 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.

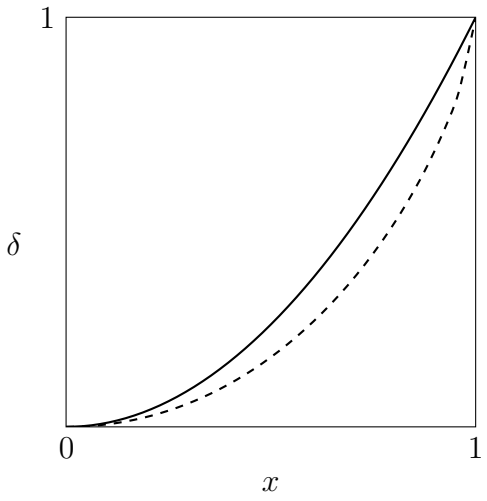
[27] In addition to modeling the policy choice about how much consumption to sacrifice for an instantaneous reduction to the hazard rate, an earlier version of this paper models the technology path as directed by policy as well. The growth model is semi-endogenous, so total potential technology growth 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 weakly 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, civilization traverses roughly the same technology path but more quickly. When future population growth is anticipated to be faster, the value of the future is higher (due to faster future technological development even if larger populations are not valued more intrinsically), so optimal policy shifts the technology path in a safer direction.

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

$$\delta_t = \bar{\delta} A_t^{\alpha} x_t^{\beta} \frac{1 - (1 - x_t)^{\epsilon}}{x_t}, \quad \epsilon \in \left(\frac{1}{2}, 1\right), \tag{57}$$

where the conditions on parameters other than ϵ are as in (5). 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). However, the transition dynamics are 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 risk and policy are illustrated below for $\epsilon = 0.6$, $A_0 = 2.03$, and otherwise the same parameter values as in Table 1.[28]

[28] 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.

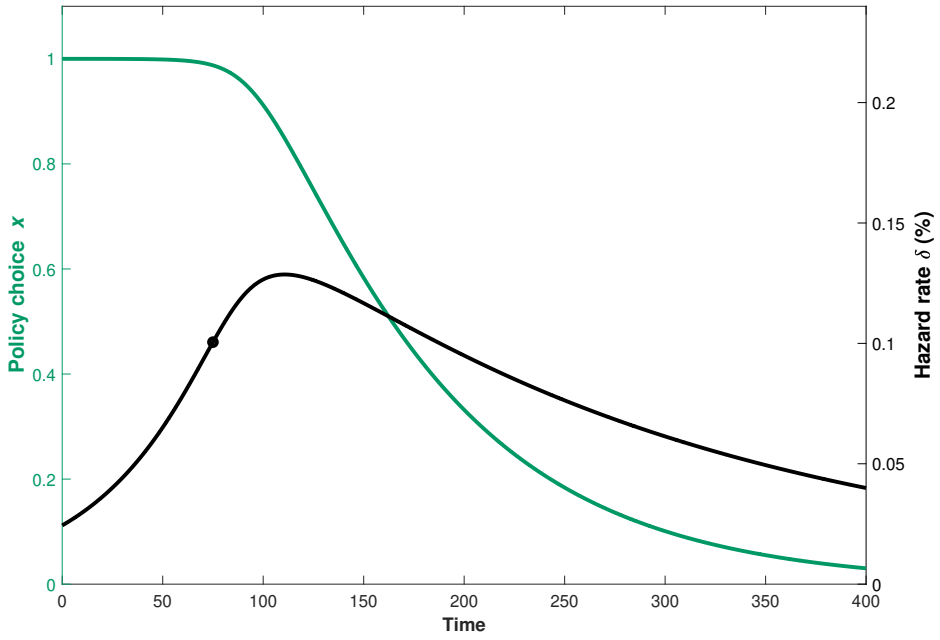


Figure B1: Evolution of the policy choice and the hazard rate along the optimal path given a lower Inada condition on safety expenditure

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

B.4.2 Safety in redundancy

The constant elasticity hazard function of Sections 3.2–3.4, 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 of survival through a given period is the probability that all C consumption units, 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

$$\delta(A_t, x_t) = 1 - \left(1 - p\tilde{b}^{\frac{1-x_t}{x_t}}\right)^{A_t x_t}, \quad \tilde{b} \in (0,1). \tag{58}$$

The continuous-time analog to (58) is

$$\delta(A_t, x_t) = A_t x_t e^{-b\frac{1-x_t}{x_t}}, \quad b > 0 \tag{59}$$

(see Appendix C.6.1).

Since hazard function (59) 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 4–6 tell us that (59) yields a Kuznets curve, with δ eventually falling quickly enough to permit survival.

Proposition 8. ***Long-run policy and risk given safety in redundancy*** *Given hazard function (59), the optimal path features*

$$\lim_{t\to\infty} x_t t = \frac{b}{g\gamma}, \tag{60}$$

$$\lim_{t\to\infty} g_{\delta t} = -g(\gamma - 1). \tag{61}$$

Proof. See Appendix C.6.2. □

Thus the decline in policy choice here is slower than in the constant elasticity case: x declines proportionally to $1/t$, not exponentially. This is because a redundancy-based model 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 (61) 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 original model. 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:

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

Intuitively it is because, in a redundancy-based model, smaller consumption sacrifices (linear rather than proportional) are needed for 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 (60), and partially in the form of faster declines in the hazard rate, as described by (61).

B.5 Transition risk: Optimal technology growth

B.5.1 Without policy, optimality of stagnation given $\zeta = 1$

Suppose first that $\zeta = 1$ and $\alpha = -1$, so that

$$\delta_t = \bar{\delta}\frac{\dot{A}_t}{A_t}.$$

As noted in the body text, this model is precisely the Russian roulette model of Jones (2016), with $\bar{\delta}$ representing the variable there denoted π.

Jones finds in his setting that, with $\gamma > 1$, it is optimal for technology to grow only to a finite level $\hat{A}$. In our notation, this is because stagnation at some $\hat{A}$, with no risk, yields constant flow utility of $u(\hat{A})$ and a constant value of the future of $v(\hat{A}) \equiv u(\hat{A})/\rho$. It is thus optimal to halt growth at the technology level at which the future benefits of stagnating at a slightly higher A equal the costs via temporarily inducing a positive hazard rate:

$$\begin{aligned} v'(\hat{A}) &= \frac{\partial \delta}{\partial \dot{A}} \cdot v(\hat{A}) \\ \implies \frac{u'(\hat{A})}{\rho} &= \frac{\bar{\delta}}{\hat{A}}\frac{u(\hat{A})}{\rho} \qquad (62) \\ \implies \hat{A} &= \Big(\frac{\bar{\delta}+\gamma-1}{\bar{\delta}}\Big)^{\frac{1}{\gamma-1}}. \qquad (63) \end{aligned}$$

When $\alpha = -1$, we can derive an analytic solution for the optimal technology level (63) at which to stagnate. Though this is not possible for other values of α, it is easy to verify that, for any $\alpha \geq -\gamma$, this result does not qualitatively change. Equality (62) is then modified to

$$\begin{aligned} \frac{u'(\hat{A})}{\rho} &= \bar{\delta}\hat{A}^{\alpha}\frac{u(\hat{A})}{\rho} \\ \implies \hat{A}^{-(\alpha+\gamma)} &= \bar{\delta}u(\hat{A}). \qquad (64) \end{aligned}$$

Given $\alpha + \gamma > 0$, the left-hand side falls strictly monotonically from 1 to 0 as $\hat{A}$ rises from 1 to ∞. The right-hand side rises strictly monotonically from 0 to $\bar{\delta}/(\gamma-1) > 0$ as $\hat{A}$ rises from 1 to ∞. There is thus a unique $\hat{A} > 1$ at which (64) is satisfied: that is, at which technology growth is preferred to stagnation iff $A < \hat{A}$.

B.5.2 Without policy, no optimal stagnation given $\zeta \neq 1$

If we further generalize from $\zeta = 1$ to arbitrary ζ, however, we find that the result that stagnation is optimal when $\zeta = 1$ is knife-edge.

Let $v_t(A_{(\cdot)})$ denote the value of the future at $t \geq 0$ given technology path $A_{(\cdot)}$. As baseline, choose a technology path $A_{(\cdot)}$ satisfying A1 and A2.

If $\zeta < 1$, then at every t, and for every technology level $\overline{A} > A_t$, there is a differentiable and weakly increasing technology path $\tilde{A}_{(\cdot)}$ with $\tilde{A}_s = A_s$ for all $s \leq t$, $\tilde{A}_{\bar{t}} = \overline{A}$ for some $\bar{t} > t$, and $v_t(\tilde{A}_{(\cdot)}) > v_t(A_{(\cdot)})$.

To construct such a path, choose t and $\overline{A} > A_t$. Observe that, if $\dot{\tilde{A}}_A$ equals a constant value $\dot{\tilde{A}}$ for $A \in (A_t, \overline{A})$, the cumulative risk endured on path $\tilde{A}_{(\cdot)}$ from A_t to $\overline{A}$ equals

$$\int_{A_t}^{\overline{A}} \bar{\delta} A^{\alpha} \dot{\tilde{A}}^{\zeta - 1} dA,$$

which $\to 0$ as $\dot{\tilde{A}} \to \infty$. With $\zeta < 1$, therefore, sufficiently rapid growth from A_t to $\overline{A}$ approximates an immediate, risk-free jump from A_t to $\overline{A}$, as in the state risk "$\zeta = 0$" case.

Now let

$$\underline{t} \equiv \min\{t : \tilde{A}_t = \overline{A}\} = \frac{\overline{A} - A_t}{\dot{\tilde{A}}},$$
$$\bar{t} \equiv \sup\{t : A_t < \overline{A}\},$$

noting that $\bar{t}$ may be infinite, and choose $\overline{A}$ and $\dot{\tilde{A}}$ so that $\dot{\tilde{A}} > \dot{A}_s$ for all $s \in [t, \underline{t})$. This is possible for some sufficiently high $\dot{\tilde{A}}$ by the right-continuous differentiability of $A_{(\cdot)}$, and ensures that $\tilde{A}_s > A_s$ throughout this interval. Suppose that $\tilde{A}_t = \overline{A}$ for $t \in [\underline{t}, \bar{t}]$ and $\tilde{A}_t = A_t$ for $t > \bar{t}$—i.e. that the new path halts growth at $\overline{A}$ until the old path has caught up, if ever, after which the paths are identical. Then $\tilde{A}_{(\cdot)}$ offers strictly higher consumption than $A_{(\cdot)}$ across $(t, \bar{t})$ in exchange for arbitrary little up-front risk and no subsequent increases in the hazard rate.

Incidentally, this framework makes clear that, in the absence of any costs to technological development besides transitional existential risk, with $\zeta < 1$ there is no optimal continuous technology path. An immediate jump in the technology level is always desirable, and a larger jump is always preferable to a smaller one. Furthermore, if one introduces R&D costs to the model, an optimal path will exist only if the costs are sufficiently convex in the speed of technological development. Otherwise, attempts to identify an optimal technology path will encounter the "chattering" problem: rapid alternations between slow and fast growth will be preferred to continuous growth, because they can achieve a given quantity of technological progress over a given interval of time while contributing less to cumulative risk.

Stagnation is not optimal given $\zeta < 1$ because, due to the "upper Inada condition" on $\delta \propto \dot{A}^{\zeta}$ with $\zeta < 1$, sufficiently fast technological development carries arbitrarily little

risk per unit of new technology. Stagnation is not optimal given $\zeta > 1$ because, since $\lim_{\dot{A}\to 0} \frac{\partial \delta}{\partial \dot{A}} = 0$ when $\delta \propto \dot{A}^{\zeta}$ with $\zeta > 1$, sufficiently slow technological development carries arbitrarily little risk per unit of new technology.

To see this, consider the optimal technology growth rate at t given a technology path $A_{(\cdot)}$ with $A_t = \hat{A} > 1$ and $\dot{A}_s = 0$ for $s > t$. Unlike in the $\zeta < 1$ case, there is an optimal technology growth rate to adopt at t: the rate $\dot{A}^*$ that sets the marginal expected utility benefit (via increased future consumption) of marginally increasing $\dot{A}$, per unit time that $\dot{A}$ is increased, equal to the marginal expected utility cost per unit time (via an increased hazard rate at t):

$$v'(\hat{A}) = \bar{\delta}\hat{A}^{\alpha}\zeta\dot{A}^{*\zeta-1}v(\hat{A})$$
$$\implies \dot{A}^* = \left(\frac{\gamma-1}{\bar{\delta}} \cdot \frac{\hat{A}^{-(\alpha+\gamma)}}{1-\hat{A}^{1-\gamma}}\right)^{\frac{1}{\zeta-1}} > 0.$$

Likewise, given a technology path $A_{(\cdot)}$ with $\lim_{t\to\infty} A_t = \hat{A} < \infty$, the optimal technology growth rate must satisfy the equality above in the limit. Since $A_{(\cdot)}$ cannot approach a finite upper asymptote if $\dot{A}$ is bounded above zero, no such technology path is optimal.

B.5.3 With policy, analogous results for ζ-threshold $1 + \frac{\beta}{\gamma-1}$

Throughout this section we will assume hazard function (27) with $\zeta > 0$:

$$\delta_t = \delta(A_t, \dot{A}_t, x_t) = \bar{\delta}A_t^{\alpha}\dot{A}_t^{\zeta}x_t^{\beta} \quad \bar{\delta} > 0,\ \zeta > 0,\ \beta > 1.$$

For simplicity we will also assume that the baseline technology path features stagnation at technology level $\hat{A}$. We will then consider the impact per unit time of an instantaneous marginal increase to the technology growth rate $\dot{A}_t$.

We will see that, in the $\zeta < 1 + \frac{\beta}{\gamma-1}$ case, as in the $\zeta < 1$ case without policy, there is no optimal growth rate: sufficiently fast growth is always preferable to stagnation. In the $\zeta > 1 + \frac{\beta}{\gamma-1}$ case, as in the $\zeta > 1$ case without policy, growth may be "too fast", but there is still no technology level at which it is optimal to stagnate.

However, the $\zeta = 1 + \frac{\beta}{\gamma-1}$ case is *not* closely analogous to the $\zeta = 1$ case without policy. Instead, for low values of $\hat{A}$ it resembles the $\zeta < 1 + \frac{\beta}{\gamma-1}$ case, with no optimal technology growth rate, and for high values of $\hat{A}$ it resembles the $\zeta < 1 + \frac{\beta}{\gamma-1}$ case, in which slow growth is preferable both to fast growth and to stagnation. Intuitively, this is because $\zeta = 1 + \frac{\beta}{\gamma-1}$ implies $\zeta > 1$. Since slow growth without policy is preferable to stagnation given $\zeta > 1$, and since introducing the option to mitigate risk with $x_t < 1$ does not remove the option of slow growth without policy, introducing the policy option cannot render stagnation optimal.

In this setting, there are two state variables: S_t and A_t. There are two choice variables: policy x_t and the technology growth rate $\dot{A}_t$.[29] Given $S_t = 1$, the marginal net impacts on expected utility of a marginal increase in $\dot{A}_t$, per unit time, is given by the respective derivative of the Hamiltonian expression

$$u(\hat{A}, x_t) - v_t\delta(\hat{A}, \dot{A}_t, x_t) + a_t\dot{A}_t \tag{65}$$

(adapted from Appendix A.1 below), where a is the costate variable on technology.

Under the $x_t \leq 1$ constraint, the optimal choice of x_t given $\dot{A}_t$ is given by the first order conditions $\partial\mathcal{L}/\partial x_t = 0$, $\partial\mathcal{L}/\partial\mu_t \geq 0$, $\mu_t\partial\mathcal{L}/\partial\mu_t = 0$ on the Lagrangian

$$\mathcal{L} = u(\hat{A}, x_t) - v_t\delta(\hat{A}, \dot{A}_t, x_t) + a_t\dot{A}_t + \mu(1 - x_t). \tag{66}$$

This reduces to

$$x_t = \min\left(1, \left(\bar{\delta}\beta\hat{A}^{\alpha+\gamma-1}\dot{A}_t^{\zeta}v(\hat{A})\right)^{-\frac{1}{\beta+\gamma-1}}\right), \tag{67}$$

with $\mu_t > 0$ iff the second term of the above minimum—the unconstrained optimal choice of x_t—is greater than 1. (This is adapted from (37)–(38).)

To find the marginal net impact on expected utility of a marginal increase in $\dot{A}_t$ per unit time, given that x_t is set optimally in response, we can take the first derivative of (66) with respect to $\dot{A}_t$ and evaluate it at $x_t =$ (67). Because (65) and (66) are continuously differentiable in $\dot{A}_t$, x_t, and μ_t, by the envelope theorem we can differentiate (66) with respect to $\dot{A}_t$ and then substitute $x_t =$ (67), rather than accounting for the impact of changing $\dot{A}_t$ on the choice of x_t by substituting (67) into (65) and differentiating the result with respect to $\dot{A}_t$.

Finally, given technology level $A_t = \hat{A}$ and permanent stagnation after t, the value of the costate variables at t are straightforward. The value of [saving] civilization at t is $v(\hat{A})$, and the value of a marginal increase in the technology level is the value of an equal marginal increase in consumption at all future periods:

$$v_t = v(\hat{A}) = \frac{1}{\rho} \cdot \frac{\hat{A}^{1-\gamma} - 1}{1 - \gamma},$$
$$a_t = v'(\hat{A}) = \frac{\hat{A}^{-\gamma}}{\rho}.$$

The marginal net impact on expected utility of a marginal increase in $\dot{A}_t$ per unit

[29] It would be equivalent, and more standard but in this case more complex, to define a new choice variable ϕ_t such that the technology law of motion is $\dot{A}_t = \phi_t$.

time is therefore

$$\begin{aligned} d(\dot{A}_t) &\equiv \frac{\hat{A}^{-\gamma}}{\rho} - v(\hat{A})\bar{\delta}\hat{A}^{\alpha}\zeta\dot{A}_t^{\zeta-1}x_t^{\beta} \qquad (68)\\ &= \frac{\hat{A}^{-\gamma}}{\rho} - v(\hat{A})\bar{\delta}\hat{A}^{\alpha}\zeta\dot{A}_t^{\zeta-1}, \qquad \dot{A}_t < \underline{\dot{A}_t};\\ &= \frac{\hat{A}^{-\gamma}}{\rho} - \zeta\Big(\bar{\delta}^{1-\gamma}v(\hat{A})^{1-\gamma}\hat{A}^{(\beta-\alpha)(\gamma-1)}\beta^{\beta}\Big)^{-\frac{1}{\beta+\gamma-1}}\dot{A}_t^{\zeta\frac{\gamma-1}{\beta+\gamma-1}-1}, \quad \dot{A}_t \geq \underline{\dot{A}_t}, \end{aligned}$$

where

$$\underline{\dot{A}_t} \equiv \left(\bar{\delta}\beta\hat{A}^{\alpha+\gamma-1}v(\hat{A})\right)^{-\frac{1}{\zeta}}$$

is the maximum growth rate at which it is optimal to set $x_t = 1$, and $v(\hat{A})$ is as defined above.

If $\zeta < 1 + \frac{\beta}{\gamma-1}$, then the exponent on $\dot{A}_t$ in (68) is negative for $\dot{A}_t \geq \underline{\dot{A}_t}$, so

$$\lim_{\dot{A}_t\to\infty} d(\dot{A}_t) = \hat{A}^{-\gamma}/\rho > 0.$$

As in the $\zeta < 1$ case without policy, this guarantees that sufficiently fast technology growth is always preferable to stagnation.

If $\zeta > 1 + \frac{\beta}{\gamma-1}$, then the exponent on $\dot{A}_t$ in (68) is always positive. There is thus a unique and positive value of $\dot{A}_t$ that sets $d(\dot{A}_t) = 0$, and this is the optimal choice of $\dot{A}_t$. Sufficiently slow technology growth is always preferable to stagnation.

If $\zeta = 1 + \frac{\beta}{\gamma-1}$, then the exponent on $\dot{A}_t$ in (68) is positive for $\dot{A}_t < \underline{\dot{A}_t}$ and zero for $\dot{A}_t \geq \underline{\dot{A}_t}$. So if $d(\underline{\dot{A}_t}) > 0$, there is no optimal growth rate: from the $A_t = \hat{A}$ margin, it is desirable, albeit perhaps briefly, to have technology grow as quickly as possible. If $d(\underline{\dot{A}_t}) < 0$, there is a unique value of $\dot{A}_t$ that sets $d(\dot{A}_t) = 0$, it lies in $(0, \underline{\dot{A}_t})$, and it is optimal.

Technically, if $d(\underline{\dot{A}_t}) = 0$, then any $\dot{A}_t \geq \underline{\dot{A}_t}$ is optimal at $A_t = \hat{A}$; but once $A_t > \hat{A}$, we will have $d(\underline{\dot{A}_t}) < 0$, and a unique optimal growth rate which is positive but finite.

C Supplemental proofs

C.1 Proof of Proposition 3

The proof is similar to the proof of Proposition 7a (Appendix C.5.1), which generalizes Proposition 2. As there, $v_{\underline{A}+\epsilon}$ is continuous in ϵ and $\tilde{v}_{\underline{A}+\epsilon}[\epsilon] = v_{\underline{A}+\epsilon}$ for all ϵ. In this setting, however, we cannot assume that $\tilde{v}_A[\epsilon]$ weakly increases in A or that $\tilde{v}_{\underline{A}}[\epsilon] \geq v_{\underline{A}}$ for all ϵ. We will therefore use a different strategy to uniformly bound $\tilde{v}_A[\epsilon]$, for $A \in [\underline{A}, \underline{A} + \epsilon]$, in an interval whose maximum and minimum converge to $v_{\underline{A}}$ as $\epsilon \to 0$.

Let $\underline{t}$ denote the time at which $A_t = \underline{A}$. An acceleration $\tilde{A}_{(\cdot)}$, featuring technology growth rate $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$ until technology level $\underline{A} + \epsilon$, features technology growth at rate $\dot{\tilde{A}}$ across times

$$(\underline{t}, \underline{t} + \epsilon/\dot{\tilde{A}}).$$

More generally, the acceleration path reaches technology level $A \in [\underline{A}, \underline{A} + \epsilon]$ at time

$$\tilde{t}(A) \equiv \underline{t} + (A - \underline{A})/\dot{\tilde{A}}.$$

$\tilde{v}_A[\epsilon]$ is the maximum value of survival $\tilde{v}_{\tilde{t}(A)}$, across feasible policy paths, achievable at $\tilde{t}(A)$ given technology path $\tilde{A}_{(\cdot)}[\epsilon]$. It can thus be lower-bounded by one such achievable value of survival, such as that achieved given $x_t = 1$ for $t \in [\tilde{t}(A), \underline{t} + \epsilon/\dot{\tilde{A}})$. Since $\tilde{A}_t > 1$ throughout this interval, this lower bound is in turn strictly greater than the value of survival at $\tilde{t}(A)$ given *no* flow utility enjoyed throughout the interval.

Remembering that $\tilde{v}_{\underline{A}+\epsilon}[\epsilon] = v_{\underline{A}+\epsilon} > 0$ for any ϵ, we thus have

$$\begin{aligned}
\tilde{v}_A[\epsilon] &\geq \int_{\tilde{t}(A)}^{\underline{t}+\epsilon/\dot{\tilde{A}}} e^{-\rho(t-\tilde{t}(A))} e^{-\int_{\tilde{t}(A)}^{t} \bar{\delta} \tilde{A}_s^{\alpha} \dot{\tilde{A}}^{\zeta} ds} u(\tilde{A}_t) dt \\
&\quad + e^{-\rho(\underline{t}+\epsilon/\dot{\tilde{A}}-\tilde{t}(A))} e^{-\int_{\tilde{t}(A)}^{\underline{t}+\epsilon/\dot{\tilde{A}}} \bar{\delta} \tilde{A}_s^{\alpha} \dot{\tilde{A}}^{\zeta} ds} v_{\underline{A}+\epsilon} \\
&> \underline{v}_A[\epsilon] \equiv e^{-\rho(\underline{t}+\epsilon/\dot{\tilde{A}}-\tilde{t}(A))} e^{-\int_{\tilde{t}(A)}^{\underline{t}+\epsilon/\dot{\tilde{A}}} \bar{\delta} \tilde{A}_s^{\alpha} \dot{\tilde{A}}^{\zeta} ds} v_{\underline{A}+\epsilon}. \qquad (69)
\end{aligned}$$

Because $\tilde{t}(A)$ increases in A, $\underline{v}_A[\epsilon]$ increases in A, so $\underline{v}_A[\epsilon] \geq \underline{v}_{\underline{A}}[\epsilon]$ for all $A \in [\underline{A}, \underline{A}+\epsilon]$.

$\tilde{v}_A[\epsilon]$ can be upper-bounded by the (infeasible) value of survival achieved at $\tilde{t}(A)$ given that, at $t \in [\tilde{t}(A), \underline{t} + \epsilon/\dot{\tilde{A}})$, flow utility equals its supremum of $1/(\gamma - 1)$ and the hazard rate equals 0:

$$\begin{aligned}
\tilde{v}_A[\epsilon] &< \frac{1}{\gamma - 1} \int_{\tilde{t}(A)}^{\underline{t}+\epsilon/\dot{\tilde{A}}} e^{-\rho(t-\tilde{t}(A))} dt + e^{-\rho(\underline{t}+\epsilon/\dot{\tilde{A}}-\tilde{t}(A))} v_{\underline{A}+\epsilon} \\
&< \overline{v}_A[\epsilon] \equiv \frac{1}{\gamma - 1} \int_{\tilde{t}(A)}^{\underline{t}+\epsilon/\dot{\tilde{A}}} e^{-\rho(t-\tilde{t}(A))} dt + v_{\underline{A}+\epsilon}. \qquad (70)
\end{aligned}$$

Because $\tilde{t}(A)$ increases in A, $\underline{v}_A[\epsilon]$ decreases in A, so $\underline{v}_A[\epsilon] \geq \underline{v}_{\underline{A}+\epsilon}[\epsilon]$ for all $A \in [\underline{A}, \underline{A} + \epsilon]$.

From (69), (70), the continuity of $v_{\underline{A}+\epsilon}$ in ϵ, and the fact that $\tilde{v}_{\underline{A}+\epsilon}[\epsilon] = v_{\underline{A}+\epsilon}$ for all ϵ,

$$\lim_{\epsilon \to 0} \underline{v}_{\underline{A}}[\epsilon] = \lim_{\epsilon \to 0} \overline{v}_{\underline{A}+\epsilon}[\epsilon] = v_{\underline{A}}.$$

The proof then proceeds along the lines of the proof of Proposition 7 after (91), with

$$\tilde{x}_A[\epsilon] = \min\left(1, \left(\bar{\delta}\beta A^{\alpha+\gamma-1} \dot{\tilde{A}}^{\zeta} \tilde{v}_A[\epsilon]\right)^{-\frac{1}{\beta+\gamma-1}}\right)$$

in place of (92), ultimately yielding

$$\Delta_{\underline{A},\dot{\tilde{A}}} = \delta(\underline{A}, \dot{\tilde{A}}, \tilde{x}_{\underline{A}})\dot{\tilde{A}}^{-1} - \delta(\underline{A}, \dot{A}_{\underline{A}}, x_{\underline{A}})\dot{A}_{\underline{A}}^{-1}, \tag{71}$$

where $\tilde{x}_{\underline{A}}$ is given by (29), at $A = \underline{A}$, with $\dot{\tilde{A}}$ in place of $\dot{A}_{\underline{A}}$.

If $\underline{A} \geq A^*$, (71) reduces to

$$\big(\bar{\delta}^{1-\gamma}\beta^{\beta}\underline{A}^{(\beta-\alpha)(\gamma-1)}v_{\underline{A}}^{\beta}\big)^{-\frac{1}{\beta+\gamma-1}}\Big(\dot{\tilde{A}}^{\zeta\frac{\gamma-1}{\beta+\gamma-1}-1} - \dot{A}_{\underline{A}}^{\zeta\frac{\gamma-1}{\beta+\gamma-1}-1}\Big).$$

Since $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$, this is negative if $\zeta < 1 + \frac{\beta}{\gamma-1}$, zero if $\zeta = 1 + \frac{\beta}{\gamma-1}$, and positive if $\zeta > 1 + \frac{\beta}{\gamma-1}$.

If $\underline{A} < A^*$, so that $x_{\underline{A}} = 1$, and $\dot{\tilde{A}}$ is small enough to maintain $\tilde{x}_{\underline{A}} = 1$, then (71) reduces to

$$\bar{\delta}\underline{A}^{\alpha}\big(\dot{\tilde{A}}^{\zeta-1} - \dot{A}_{\underline{A}}^{\zeta-1}\big).$$

Since $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$, this is negative if $\zeta < 1$, zero if $\zeta = 1$, and positive if $\zeta > 1$.

C.2 Proof of Proposition 4

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

$$\lim_{n\to\infty} R(n\overline{C}) = \lim_{n\to\infty}\lim_{A\to\infty} \frac{\partial\delta}{\partial x}\big(A, \frac{n\overline{C}}{A}\big)\frac{(n\overline{C})^{\gamma}}{A\rho(\gamma-1)} \leq 1. \tag{72}$$

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

$$R(C_{t_n}) \leq R(n\overline{C})\Big(\frac{C_{t_n}}{n\overline{C}}\Big)^{\gamma} \leq R(n\overline{C})\, n^{-\gamma}\ \ \forall n, \tag{73}$$

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 (72), $R(n\overline{C})n^{-\gamma} < 1$ for sufficiently large n, so by (73) and A4, 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

$$\frac{\partial \delta}{\partial x}\big(A, \frac{\underline{C}}{A}\big) \frac{1}{\rho(\gamma-1)} > A\underline{C}^{-\gamma} \tag{74}$$

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 $\underline{C}$, (74) 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

$$\frac{\partial \delta}{\partial x}\big(A_t, \frac{C_t}{A_t}\big) \frac{1}{\rho(\gamma-1)} > A_t C_t^{-\gamma} \;\; \forall t \geq \underline{t}. \tag{75}$$

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 (75) 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 $C^* = \infty$.

C.3 Proof of Proposition 5a

C.3.1 Preliminaries

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:

$$A_t^{1-\gamma} \geq \frac{\partial \delta}{\partial x}(A_t, 1)\, v_t. \tag{76}$$

It is optimal to set $x_t < 1$ as long as (76) fails, maintaining

$$A_t^{1-\gamma} x_t^{-\gamma} = \frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t \tag{77}$$

$$\implies x_t = A_t^{\frac{1-\gamma}{\gamma}} \Big(\frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t \Big)^{-\frac{1}{\gamma}}. \tag{78}$$

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

C.3.2 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 $t < T$,

$$\begin{aligned} v_t &= \int_t^\infty e^{-\rho(s-t) - \int_t^s \delta_q dq} u(C_s) ds \\ &= \int_t^T e^{-\rho(s-t) - \int_t^s \delta_q dq} u(C_s) ds \;+\; e^{-\rho(T-t) - \int_t^T \delta_q dq} v_T. \end{aligned} \tag{79}$$

Since $u(C_s) \leq u(A_s) \leq -1$ for $s \leq T$, the first term of (79) 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

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

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.

C.3.3 Proof that $\lim_{t\to\infty} x_t = 0$ if η_A is bounded above $1-\gamma$

Generalizing (78), whether or not the $x_t \leq 1$ constraint binds we have

$$x_t \leq A_t^{\frac{1-\gamma}{\gamma}} \Big(\frac{\partial \delta}{\partial x}(A_t, x_t)\, v_t \Big)^{-\frac{1}{\gamma}}. \tag{80}$$

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 A4) $A_t \to \infty$.

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

$$A_t^{\frac{1-\gamma}{\gamma}} \big(\delta(A_t, x_t) v_t \big)^{-\frac{1}{\gamma}}. \tag{81}$$

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, (81) tends to 0 as $A \to \infty$. Finally, v_t is positive for all $t \geq 0$, because by A1 and A2 $A_t > 1$ for all $t \geq 0$ (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 (80) as $A_t \to \infty$ requires $x_t \to 0$.

C.4 Proof of Proposition 6

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

$$\lim_{t\to\infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{\overline{k}}{\gamma-1}}}{A_t} \Big) \frac{t^{\frac{\overline{k}\gamma}{\gamma-1}}}{A_t \rho(\gamma-1)} < 1. \tag{82}$$

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

$$C_{t_n} \leq t_n^{\frac{k}{\gamma-1}} \;\; \forall n \geq 1. \tag{83}$$

Observe that

$$\begin{aligned} &\lim_{n\to\infty} \frac{\partial \delta}{\partial x}\Big(A_{t_n}, \frac{t_n^{\frac{k}{\gamma-1}}}{A_{t_n}} \Big) \frac{t_n^{\frac{k\gamma}{\gamma-1}}}{A_{t_n} \rho(\gamma-1)} \\ \leq &\lim_{t\to\infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{\overline{k}}{\gamma-1}}}{A_t} \Big) \frac{t^{\frac{\overline{k}\gamma}{\gamma-1}}}{A_t \rho(\gamma-1)} \cdot t^{-\frac{\overline{k}-k}{\gamma-1}\gamma} = 0, \end{aligned} \tag{84}$$

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 (82).

By (83), (84), 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 (54) and (83).

So for some $k > 1$,

$$\exists \underline{t} \ : \ C_t > t^{\frac{k}{\gamma-1}} \ \ \forall t > \underline{t}. \tag{85}$$

So (85) holds for $k = 1$ as well.

Given (85) for some $k > 1$, we have, for some $\underline{t}$ and some $\underline{k} \in (1, k)$, that for all $t > \underline{t}$

$$\begin{aligned} & (A_t x_t)^{1-\gamma} < t^{-k} \\ \implies & \frac{\partial \delta}{\partial x}(A_t, x_t) x_t \, v_t < t^{-k} \\ \implies & \delta_t v_t < t^{-k} \\ \implies & \delta_t < t^{-\underline{k}}. \end{aligned} \tag{86}$$

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 (86) that $S_\infty > 0$.

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

$$\frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{k}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{k\gamma}{\gamma-1}}}{A_t \rho(\gamma-1)} > 1 \ \ \forall t > \underline{s}. \tag{87}$$

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

$$C_{t_n} \geq t_n^{\frac{1}{\gamma-1}} \ \ \forall n \geq 1. \tag{88}$$

Observe that

$$\begin{aligned} & \lim_{n \to \infty} \frac{\partial \delta}{\partial x}\Big(A_{t_n}, \frac{t_n^{\frac{1}{\gamma-1}}}{A_{t_n}}\Big) \frac{t_n^{\frac{\gamma}{\gamma-1}}}{A_{t_n} \rho(\gamma-1)} \\ \geq & \lim_{t \to \infty} \frac{\partial \delta}{\partial x}\Big(A_t, \frac{t^{\frac{k}{\gamma-1}}}{A_t}\Big) \frac{t^{\frac{k\gamma}{\gamma-1}}}{A_t \rho(\gamma-1)} \cdot t^{\frac{1-k}{\gamma-1}\gamma} = \infty, \end{aligned} \tag{89}$$

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 (87).

By (88), (89), 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

$$\exists \underline{t} : C_t < t^{\frac{1}{\gamma-1}} \ \ \forall t > \underline{t}. \tag{90}$$

By (90) and (54), $x_t \to 0$. 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 (90),

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

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 $S_\infty = 0$.

C.5 Proof of Proposition 7

Choose an admissible technology path $A_{(\cdot)}$ and hazard function $\delta(\cdot)$.

C.5.1 Proof of part a

Choose $\underline{A}$, $\dot{\tilde{A}}$ with $\dot{\tilde{A}} > \dot{A}_{\underline{A}}$. Define $\tilde{x}_A[\epsilon]$ as $\tilde{x}_A$ given acceleration $\tilde{A}_{(\cdot)}[\epsilon]$, etc.

v_t is weakly increasing and continuous (indeed differentiable; see Appendix A.1) in t. Since A_t is continuous, increasing, and invertible in t, v_A is continuous and weakly increasing in A. $v_{\underline{A}+\epsilon}$ is therefore continuous and weakly increasing in ϵ.

From technology level $\underline{A} + \epsilon$ onward, the technology paths, and thus the paths of both consumption and the hazard rate, are identical under $A_{(\cdot)}$ and $\tilde{A}_{(\cdot)}$. So for any

ϵ (including 0), $\tilde{v}_{\underline{A}+\epsilon}[\epsilon] = v_{\underline{A}+\epsilon}$. From this, the fact that $\tilde{v}_A[\epsilon]$ is weakly increasing in A, and the fact that $\tilde{v}_{\underline{A}}[\epsilon] \geq v_{\underline{A}}$ for all ϵ, we have that for all ϵ

$$\tilde{v}_A[\epsilon] \in [\tilde{v}_{\underline{A}}, \tilde{v}_{\underline{A}+\epsilon}] \subseteq [v_{\underline{A}}, v_{\underline{A}+\epsilon}] \quad \forall A \in [\underline{A}, \underline{A}+\epsilon]. \tag{91}$$

Then by the continuity of $v_{\underline{A}+\epsilon}$ in ϵ, for any ϵ_1 there is an $\overline{\epsilon}$ such that $|v_{\underline{A}+\epsilon} - v_{\underline{A}}| < \epsilon_1 \ \ \forall \epsilon < \overline{\epsilon}$.

Adapting (37),

$$\tilde{x}_A[\epsilon] = \min\Big(1,\, x : \frac{\partial \delta}{\partial x}(A,x)A^{\gamma-1}x^{\gamma} = \frac{1}{\tilde{v}_A[\epsilon]}\Big). \tag{92}$$

By (92) and A2, $\tilde{v}_A[\epsilon] \geq v_{\underline{A}} > 0$ for all $\epsilon \geq 0, A \in [\underline{A}, \underline{A}+\epsilon]$. By D3, the implicit function theorem, and the continuity of $\min(\cdot)$, $\tilde{x}_A[\epsilon]$ is continuous in $\tilde{v}_A[\epsilon]$. So by (91) and the sentence following it, for any ϵ_2 there is an $\overline{\epsilon}$ such that, for all $\epsilon < \overline{\epsilon}$,

$$\Big|\tilde{x}_A[\epsilon] - \min\Big(1,\, x : \frac{\partial \delta}{\partial x}(A,x)A^{\gamma-1}x^{\gamma} = \frac{1}{v_{\underline{A}}}\Big)\Big| < \epsilon_2 \quad \forall A \in [\underline{A}, \underline{A}+\epsilon].$$

Again by D3, the implicit function theorem, and the continuity of $\min(\cdot)$, the second term in the absolute value is continuous in A. So for any ϵ_3 there is an $\overline{\epsilon}$ such that, for all $\epsilon < \overline{\epsilon}$,

$$\Big|\tilde{x}_A[\epsilon] - \min\Big(1,\, x : \frac{\partial \delta}{\partial x}(\underline{A},x)\underline{A}^{\gamma-1}x^{\gamma} = \frac{1}{v_{\underline{A}}}\Big)\Big| = \big|\tilde{x}_A[\epsilon] - x_{\underline{A}}\big| < \epsilon_3 \quad \forall A \in [\underline{A}, \underline{A}+\epsilon].$$

With this uniform convergence, since

$$\tilde{X}[\epsilon] - X = \int_{\underline{A}}^{\underline{A}+\epsilon} \delta\big(A, \tilde{x}_A[\epsilon]\big)\dot{\tilde{A}}^{-1}dA - \int_{\underline{A}}^{\underline{A}+\epsilon} \delta(A, x_A)\dot{A}_A^{-1}dA,$$

since $\delta(\cdot)$ is continuous in both arguments, since x_A is continuous in A, and since $\dot{A}_A^{-1}$ is right-continuous in time and thus (by continuity and monotonicity of $A_{(\cdot)}$) in A,

$$\begin{aligned}\Delta_{\underline{A},\dot{\tilde{A}}} &\equiv \lim_{\epsilon\to 0}\frac{\tilde{X}[\epsilon] - X}{\epsilon} = \delta(\underline{A}, x_{\underline{A}})\dot{\tilde{A}}^{-1} - \delta(\underline{A}, x_{\underline{A}})\dot{A}_{\underline{A}}^{-1} \\ &= \delta_{\underline{A}}\big(\dot{\tilde{A}}^{-1} - \dot{A}_{\underline{A}}^{-1}\big).\end{aligned}$$

This proves (a).

C.5.2 Proof of part b

Let $\tilde{A}_{(\cdot)}$ be an acceleration to $A_{(\cdot)}$ from $\underline{A}$ to $\overline{A}$. By the definition of an acceleration and the definition of cumulative risk,

$$\tilde{X} = X + \int_{\underline{A}}^{\overline{A}} \Big(\delta\big(A, \tilde{x}_A\big)\dot{\tilde{A}}_A^{-1} - \delta(A, x_A)\dot{A}_A^{-1}\Big)dA. \tag{93}$$

For all $A \in [\underline{A}, \overline{A})$, we have $\tilde{v}_A \geq v_A$, and thus, by (92) (dropping the "[ϵ]" arguments) and D5, $\tilde{x}_A \leq x_A$. D1, D2, and D5 imply that $\delta(\cdot)$ weakly increases in x, so $\delta(A, \tilde{x}_A) \leq \delta(A, x_A)$. So

$$\delta\big(A, \tilde{x}_A\big)\dot{\tilde{A}}_A^{-1} - \delta(A, x_A)\dot{A}_A^{-1} \leq \Delta_{A, \dot{\tilde{A}}_A} \quad \forall A \in [\underline{A}, \overline{A}].$$

This proves (b).

C.5.3 Proof of part c

If $\overline{A} < \infty$, the integral of (93) finite. So given a technology path $A_{(\cdot)}$ for which $X = \infty$ and an acceleration to $\overline{A} < \infty$, $\tilde{X} = \infty$. This proves the first part of (c).

To prove the second part of (c), it will suffice to find a hazard function $\delta(\cdot)$ and technology path $A_{(\cdot)}$ for which $X = \infty$ and a pair of accelerations $\tilde{A}_{(\cdot)}$ to $\overline{A} = \infty$, for one of which $\tilde{X}$ is finite and for the other of which $\tilde{X}$ is infinite. We have already encountered both.

For a case of the former, consider the hazard function $\delta(A_t, x_t) = A_t x_t$, discussed following Proposition 6. As discussed there, cumulative risk given optimal policy is then infinite for any technology path eventually bounded above zero.

For a case of the latter, consider hazard function (5)—$\delta(A_t, x_t) = \bar{\delta} A_t^{\alpha} x_t^{\beta}$—with baseline technology path $A_t = (t-1)^k$ $(t \geq 0)$ and acceleration $\tilde{A}_t = (t-1)^{\tilde{k}}$ $(t \geq 0)$, where

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

To verify that this is an acceleration, $A_t = (t-1)^k \implies t = 1 + A_t^{\frac{1}{k}}$, so $\dot{A}_t = k(t-1)^{k-1} \implies \dot{A}_A = kA^{\frac{k-1}{k}}$, which increases in k given $A > 1$ (so, for $t > 0$).

As shown in (26), here $X = \infty$ and $\tilde{X} < \infty$.

C.6 Safety in redundancy

C.6.1 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

$$\left(e^{-\bar{\delta}b^{\frac{1-x_t}{x_t}}s}\right)^{A_t x_t} = e^{-\bar{\delta}b^{\frac{1-x_t}{x_t}}A_t x_t\, s}. \tag{94}$$

So the probability of a catastrophe by s given locally constant A, x equals 1-(94), 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

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

C.6.2 Proof of Proposition 8

By Appendix A.1, there is a unique optimal path. By the reasoning following (9), the optimal choice of x is 1 until the (unique) time at which

$$\frac{\partial u}{\partial x_t}(A_t, x_t) = \frac{\partial \delta}{\partial x_t}(A_t, x_t)\, v_t \tag{95}$$

at $x_t = 1$, after which the optimal choice of x_t is interior and maintains equality (95).

Differentiating the utility function and hazard function (59), we have

$$\begin{aligned} A_t^{1-\gamma}x_t^{-\gamma} &= \bar{\delta}A_t e^{-b\frac{1-x_t}{x_t}}\left(1+\frac{b}{x_t}\right)v_t \\ \implies \frac{1}{v_t} &= \bar{\delta}A_t^{\gamma}e^{-b\frac{1-x_t}{x_t}}\left(x_t^{\gamma} + bx_t^{\gamma-1}\right). \end{aligned} \tag{96}$$

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

$$x_{t_n} \leq \overline{C}/A_{t_n} \quad \forall n. \tag{97}$$

Substituting (97) into (96), and recalling that $A_{t_n} \to \infty$, this would imply that the right-hand side of (96) tends to 0 across $\{t_n\}$, and thus that it is not asymptotically positive.

From (96),

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

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 $\delta_t \to 0$, $v_t \to \bar{v}$.

Divide both sides of (96) 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

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

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 (96) 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 $-g(\gamma-1) < 0$.

D 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 (5), used throughout Sections 3.2–3.4 and used to simulate Figure 3, is the special case of hazard function (57), used to simulate Figure B1, with $\epsilon = 1$. The calculations below therefore apply to both 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

$$v_t = \frac{1}{\bar{\delta}} \frac{A_t^{1-\gamma-\alpha} x_t^{2-\gamma-\beta}}{(\beta-1)\big(1-(1-x_t)^{\epsilon}\big) + \epsilon x_t (1-x_t)^{\epsilon-1}} \tag{98}$$

$$\implies \dot{v}_t = v_t \Big((1-\gamma-\alpha)g + (2-\gamma-\beta)\frac{\dot{x}_t}{x_t} - \epsilon \frac{\beta - (\epsilon+\beta-1)x_t}{(\beta-1)(1-x_t)^{1-\epsilon} + 1 - \beta + (\epsilon+\beta-1)x_t} \frac{\dot{x}_t}{1-x_t} \Big). \tag{99}$$

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

$$\begin{aligned} \dot{v}_t &= v_t(\rho + \delta_t) - u(c_t) \\ &= v_t \Big(\rho + \bar{\delta} A_t^{\alpha} x_t^{\beta-1} (1-(1-x_t)^{\epsilon}) \Big) - \frac{(A_t x_t)^{1-\gamma} - 1}{1-\gamma}. \end{aligned} \tag{100}$$

Substituting (98) into (99) and (100), setting the results equal, and solving for $\dot{x}_t$ yields

$$\begin{aligned} \dot{x}_t = {} & x_t \big((\beta-1)(1-x_t)^{1-\epsilon} + 1 - \beta + (\epsilon+\beta-1)x_t \big)(1-x_t) \\ & \Big((2-\gamma-\beta)\big((\beta-1)(1-x_t)^{1-\epsilon} + 1 - \beta \\ & + (\epsilon+\beta-1)x_t\big)(1-x_t) - \epsilon(\beta - (\epsilon+\beta-1)x_t)x_t \Big)^{-1} \\ & \Big(\rho + \bar{\delta} A_t^{\alpha} x_t^{\beta-1}(1-(1-x_t)^{\epsilon}) - g(1-\alpha-\gamma) - \\ & \frac{(A_t x_t)^{1-\gamma} - 1}{1-\gamma} \bar{\delta} A_t^{\alpha+\gamma-1} x_t^{\beta+\gamma-2} \big((\beta-1)(1-(1-x_t)^{\epsilon}) + \epsilon x_t (1-x_t)^{\epsilon-1} \big) \Big). \end{aligned} \tag{101}$$

Differentiating the hazard function (57) with respect to t yields

$$\dot{\delta}_t = \bar{\delta} A_t^{\alpha} x_t^{\beta} \frac{1-(1-x_t)^{\epsilon}}{x_t} \Big(\alpha g + (\beta-1)\frac{\dot{x}_t}{x_t} + \epsilon \frac{(1-x_t)^{\epsilon}}{1-(1-x_t)^{\epsilon}} \frac{\dot{x}_t}{1-x_t} \Big). \tag{102}$$

Scripts for replicating Figures 3 and B1 using (101) and (102), and the estimate of S_∞ following Figure 3, are provided here: https://philiptrammell.com/static/ERAG_code.zip.