Existential Risk and Growth

Leopold Aschenbrenner*
Columbia University

September 2, 2019 – Version 0.5
Preliminary

CLICK HERE FOR THE MOST RECENT VERSION

Abstract

Technological innovation can create or mitigate risks of catastrophes—such as nuclear war, extreme climate change, or powerful artificial intelligence run amok—that could imperil human civilization. What is the relationship between economic growth and these existential risks? In a model of endogenous and directed technical change, with moderate parameters, existential risk follows a Kuznets-style inverted U-shape. This suggests we could be living in a unique "time of perils," having developed technologies advanced enough to threaten our permanent destruction, but not having grown wealthy enough yet to be willing to spend much on safety. Accelerating growth during this "time of perils" initially increases risk, but improves the chances of humanity's survival in the long run. Conversely, even short-term stagnation could substantially curtail the future of humanity. Nevertheless, if the scale effect of existential risk is large and the returns to research diminish rapidly, it may be impossible to avert an eventual existential catastrophe.

*Contact: leopold.aschenbrenner@columbia.edu. I am very grateful to Philip Trammell for his generous mentorship and help with this project. I am also grateful to the Centre for Effective Altruism and the University of Oxford's Future of Humanity Institute for their support.

1 Introduction

The last two centuries of technological innovation have brought enormous prosperity. Yet some of those technological innovations have also created the possibility of catastrophes such as nuclear war, extreme climate change, or bioengineered pandemics. In the next century, powerful artificial intelligence (AI) might engender amazing advances, but some worry that it could run amok. Increasing attention is being paid to these so-called "existential" risks that could imperil human civilization. In particular, existential risks are those that threaten human extinction or could otherwise irreversibly curtail the potential of humankind (such as a war that permanently sends us back to the Stone Age); see Bostrom (2002), Posner (2004), and Farquhar et al. (2017). Some philosophers argue that because of the potential "astronomical" value associated with the long-run future of humanity, mitigating existential risk should be of paramount concern; see Bostrom (2003).

However, most people generally have a positive rate of pure time preference; they do not care much about the long-run future of humanity. What happens to existential risk when resources are allocated impatiently? In particular, what is the interaction between economic growth and existential risk? Does faster economic growth accelerate the development of dangerous new technologies, thereby increasing the probability of an existential catastrophe?

I develop a model of endogenous and directed technical change, involving a tradeoff between consumption and safety. Consumption and the associated technologies carry some risk of disaster, which can be mitigated by spending on safety and developing safety technology. The outcome turns out to critically depend on the scale effect of existential risk—that is, how proportionally growing both consumption and safety affects existential risk. If existential risk decreases with scale, no special concern for safety is required for risk to fall to zero exponentially. If existential risk increases with scale moderately, the level of existential risk may follow an inverted U-shape. This grounds the intuition of some prominent thinkers, like Sagan (1994) and Parfit (2011), that human civilization could be passing through a unique "time of perils." We may have advanced enough to create technologies that threaten our permanent destruction, but not yet grown wealthy enough to be willing to spend much on safety. During this "time of perils," accelerating growth initially increases risk, but perhaps counterintuitively improves the chances of humanity's survival in the long run. Conversely, even short-term stagnation substantially hurts the changes of humanity's survival in the long run.

Finally, if the scale effect of existential risk is too large and the returns to research diminish too rapidly, it is impossible to avert an eventual existential catastrophe.

This paper relates to the literature on the relationship between economic growth and environmental degradation (see Brock and Taylor (2005) for an overview). Most relevant is Stokey (1998), demonstrating that if the marginal utility of consumption diminishes rapidly, there is an inverted U-shape relationship between economic development and pollution; this relationship is often called the "environmental Kuznets curve." I find that the level of existential risk may follow a similar inverted U-shape. However, Stokey (1998) looks at environmental degradation that additively reduces utility; existential risk that imperils the survival of human civilization is a quite different concern.

To model people's concern about risks of existential catastrophes, I build on previous work on the value of life. Hall and Jones (2007) show that for a large class of conventional preferences, as consumption grows and the marginal utility of consumption declines, it becomes relatively more valuable to purchase additional days of life rather than increasing consumption on any given day of life. Jones (2016) shows how this can lead society to value safety over consumption growth, resulting in optimal consumption growth lower than what is feasible. In his richer endogenous growth model, lifesaving goods can be purchased to increase people's lifespan. I build on this model to look at existential risk. Unlike the mortality in Jones's model, I model existential risk as increasing in consumption spending. Moreover, I model existential risk as depending on total consumption and total safety spending, not on per-capita variables as in Jones's model.

Critically, modeling existential risk as depending on total instead of per-capita variables allows for a potential scale effect. Previous work has made the implicit assumption that the risk of catastrophe stays constant with scale. In particular, Martin and Pindyck (2015, 2019) and Aurland-Bredesen (2019) posit a fixed set of possible catastrophes, which would each require a constant permanent tax on consumption to avert. Yet this is a knife-edge assumption: holding safety spending constant as a fraction of output only holds risk constant when the scale effect is exactly zero. This paper generalizes from this knife-edge assumption, illustrating the divergent dynamics of the cases when existential risk decreases, increases, or increases very rapidly with scale.

The rest of this paper is organized as follows. Section 2 presents the

economic environment of the model and a benchmark "rule of thumb allocation." Section 3 presents the asymptotic (impatient) optimal growth path, highlighting how the scale effect of existential risk matters for the long run. Section 4 discusses empirical evidence on the model parameters. Section 5 illustrates the transition path of the case with a moderate scale effect, yielding the inverted U-shape path of existential risk. Section 6 analyzes what happens to existential risk when growth accelerates. Section 7 presents conclusions.

2 The Economic Environment

I look at an endogenous idea-based growth model based on the Jones (1995) version of the Romer (1990) model. Similar to Jones (2016), this model features a consumption and a safety sector with directed technical change (see also Acemoglu (2002) and Acemoglu et al. (2012)).

2.1 Setup

The economy features a consumption sector, producing consumption good C_t, and a safety sector, producing safety good H_t. Total production in each sector is given by:

$$C_t = \left[\int_0^{A_t} x_{it}^{1/(1+\alpha)} di\right]^{1+\alpha} \quad \text{and} \quad H_t = \left[\int_0^{B_t} z_{it}^{1/(1+\alpha)} di\right]^{1+\alpha}. \tag{1}$$

Each sector uses a variety of intermediate goods to produce output. A different set of ideas is used for each sector: A_t represents consumption technologies, while B_t represents safety technologies. The safety goods z_{it} are goods that reduce the risk of catastrophe, from pollution mitigation to nuclear nonproliferation treaties to laboratory protection that prevents the outbreak of a bioengineered pandemic.

Once a technological variety has been discovered, one unit of labor can be used to produce one unit of that variety. The total number of workers is denoted L_t, so the resource constraint for workers is

$$L_{ct} + L_{ht} \leq L_t \quad \text{given} \quad L_{ct} \equiv \int_0^{A_i} x_{it} di,\ L_{ht} \equiv \int_0^{B_t} z_{it} di. \tag{2}$$

People can also produce ideas. These people are called scientists, and the production functions for ideas are given by

$$\dot{A}_t = S_{at}^{\lambda} A_t^{\phi} \quad \text{and} \quad \dot{B}_t = S_{bt}^{\lambda} B_t^{\phi}, \tag{3}$$

where as in Jones (1995), I assume $\phi < 1$ and $0 < \lambda \leq 1$.

Our resource constraints are then

$$S_{at} + S_{bt} \leq S_t \quad \text{and} \quad S_t + L_t \leq N_t, \tag{4}$$

where N_t represents the total number of people. People can become either workers or scientists, and workers and scientists can in turn either work in the consumption sector or work in the safety sector.

Next, consider existential risk. Jones (2016) considers individual-level mortality that can be reduced with lifesaving goods. I instead wish to consider risks that threaten the survival of humanity as a whole.

These risks differ from Jones's individual-level mortality in two critical ways. First, these risks are man-made: without the technological development that enables modern levels of consumption, we would not have to fear catastrophic climate change, dangerous AI, or nuclear war. Risk may increase with more consumption: higher consumption may mean more carbon emissions, more powerful AI engineered, or more potentially dangerous biotech. At the same time, the risk of an existential catastrophe may be mitigated by investing in safety: we can move to a lower-emission energy system or engineer more reliable nuclear weapon locks ("permissive action links") to reduce the risk of accidental nuclear war. Thus, unlike in Jones (2016), where mortality is only a function of spending in the lifesaving sector, existential risk in this model is a function of both consumption and safety spending. Growth in consumption is thus not purely positive, but creates risks. This model formalizes the idea of "differential technological development," as articulated by Bostrom (2002): existential risk depends on the relative rate of development of potentially dangerous technologies versus technologies that ameliorate these hazards.

The second crucial difference to Jones (2016) is that existential risks depend on *total* consumption and *total* safety spending—not on per-capita consumption and per-capita safety spending. The risk of catastrophic climate change depends on total emissions, not per-capita emissions; the risk of a bioengineered superbug escaping depends on the total amount of hazardous biotech, not on per-capita hazardous biotech; the risk of a nuclear winter

depends on the total number of nuclear weapons, not per-capita nuclear weapons. Similarly, existential risk mitigation depends on total spending on climate change abatement, biosecurity, AI safety, etc. This introduces a scale effect: risk depends on the total size of the economy, similar to how technological development depends on scale in endogenous idea-based growth models.

I will assume that an existential catastrophe results in permanent zero utility thereafter. This assumption should be exactly valid in the case of human extinction. It should also be a valid approximation for most existential catastrophes that, while not quite killing everybody, irreversibly curtail the potential of humankind (such as a war that permanently sends us back to the Stone Age).[1]

Mathematically, human civilization face a time-varying hazard rate δ_t. This represents a stochastic probability of an existential catastrophe. The probability that human civilization survives to date t (starting from date 0) is given by

$$M_t = e^{-\int_0^t \delta_s ds}, \tag{5}$$

corresponding to the laws of motion

$$\dot{M}_t = -\delta_t M_t, \quad M_0 = 1. \tag{6}$$

The hazard rate is endogenous, and as explained above increases with total consumption and decreases with total safety spending:

$$\delta_t = \overline{\delta} C_t^{\epsilon} H_t^{-\beta}. \tag{7}$$

For those concerned about the long-run future of humanity, the key variable is $M_\infty = \lim_{t\to\infty} M_t = e^{-\int_0^\infty \delta_s ds}$. This represents the probability that human civilization does *not* succumb to an existential catastrophe and enjoys a long future with astronomical value.[2] Critically, note that M_∞ is only greater than zero iff $\int_0^\infty \delta_s ds$ is bounded.

[1]Some have considered other risks, such as a creeping, irreversible spread of global authoritarianism; see Caplan (2008). Such a risk would not be covered by this model. I wish to focus on catastrophes that would kill most of the people alive at the time or make people's lives so miserable so as to reduce their utility to roughly zero.

[2]Note that surviving to time "infinity" in this model does not literally mean human civilization survives forever. Instead, it means human civilization does not destroy itself; there are other natural/physical limits to the survival of human civilization that are not considered here. In particular, there might be natural sources of extinction risk, but Snyder-Beattie et al. (2019) find that these are negligible compared to potential anthropogenic extinction risks. Thus, I focus on anthropogenic existential risk in this paper.

Given $c_t \equiv C_t/N_t$ and $h_t \equiv H_t/N_t$, expected lifetime utility for a representative agent is

$$U = \int_0^\infty e^{-\rho t} u(c_t) M_t dt, \tag{8}$$

where flow utility is isoelastic in consumption:

$$u(c_t) = \overline{u} + \frac{c_t^{1-\gamma}}{1-\gamma}. \tag{9}$$

The parameter $\overline{u}$ is a constant that specifies the upper bound of the utility of life relative to death (with the utility of death implicitly normalized to 0) in the case where $\gamma > 1$ and thus $c^{1-\gamma}/(1-\gamma)$ is negative. See Hall and Jones (2007) for a discussion of this constant.

Finally, I assume an exogenous positive rate of population growth:

$$\dot{N}_t = \overline{n} N_t. \tag{10}$$

Given the symmetry of our setup, the x_{it} and z_{it} can simply be allocated symmetrically across varieties. I will impose this throughout the rest of the paper.

There are then three allocative decisions that need to be made: the fraction of total scientists working on consumption (vs. safety), the fraction of total workers working on consumption (vs. safety), and the fraction of the population that is a scientist (vs. a worker). These three allocative decisions can be represented by three variables:

$$1.\ s_t \equiv \frac{S_{at}}{S_t} \tag{11}$$

$$2.\ \ell_t \equiv \frac{L_{ct}}{L_t} \tag{12}$$

$$3.\ \sigma_t \equiv \frac{S_t}{N_t} \tag{13}$$

2.2 Rule of Thumb Allocation

As a benchmark, it will be helpful to consider a simple "rule of thumb" allocation, as in Jones (2016). This rule of thumb allocation is analogous to Solow's (1956) assumption of a fixed saving rate in his version of the neoclassical growth model. In particular, I will consider a rule of thumb

allocation where the fraction of scientists and labor working on safety is fixed. Later, I will consider the optimal allocation, in which the fraction of resources dedicated to safety can evolve.

Proposition 1. ***Balanced growth under rule of thumb allocation*** *Consider a rule of thumb allocation where $s_t = \overline{s}$, $\ell_t = \overline{\ell}$, and $\sigma_t = \overline{\sigma}$, all strictly between zero and one. There exists a balanced growth path such that*

$$g_A^* = g_B^* = \frac{\lambda \overline{n}}{1-\phi}, \tag{14}$$

$$g_c^* = g_h^* = \overline{g} \equiv \frac{\alpha \lambda \overline{n}}{1-\phi}, \tag{15}$$

$$g_\delta^* = (\epsilon - \beta)(\overline{g} + \overline{n}), \tag{16}$$

with

$$\delta_t \to 0 \text{ if } \epsilon < \beta, \qquad \delta_t \to \delta^* > 0 \text{ if } \epsilon = \beta, \qquad \delta_t \to \infty \text{ if } \epsilon > \beta. \tag{17}$$

Proof. See Appendix A.1. □

Given our symmetric production functions and fixed allocation of labor and scientists to safety and consumption, the long-run growth of both sectors looks like growth in the standard Jones (1995) version of the Romer (1990) model. In particular, given the diminishing returns to research and the non-rivalry of ideas, the long-run growth of both consumption and safety depends on population growth. Moreover, given that a fixed fraction of workers and scientists is allocated to the safety sector, both safety and consumption per capita grow at the same rate.

What is important to note about this rule of thumb allocation is what happens to existential risk. In the case that $\epsilon < \beta$, i.e. safety is more potent in reducing risk than consumption is in increasing it, the hazard rate δ falls to zero at an exponential rate. Therefore, $\int_0^\infty \delta_s ds$ is bounded, which implies that the long-run probability of human civilization's survival, M_∞, is strictly greater than zero. In the knife-edge case of $\epsilon = \beta$, the hazard rate converges to a constant, implying $M_\infty = 0$. In the case that $\epsilon > \beta$, i.e. consumption is more potent in increasing risk than safety is in decreasing it, the hazard rate increases exponentially. This causes not only $M_\infty = 0$, but in fact $\delta \to \infty$, so the instantaneous probability of an existential catastrophe approaches 1.

Here, we begin to see the central role of $\epsilon - \beta$. Recall that $\delta_t = \overline{\delta} C_t^\epsilon H_t^{-\beta}$. Thus, $\epsilon - \beta$ represents the scale effect of existential risk. If $\epsilon < \beta$, risk

decreases with scale. Then, the future of humanity is bright: even if the allocation of resources to safety stays fixed, existential risk decreases exponentially. However, if $\epsilon > \beta$, existential risk increases with scale. Then, more scale doesn't lead to more nonrival ideas and thereby more output—as it does in the classic Romer/Jones endogenous idea-based growth model—but more scale also increases risk. In the rule of thumb allocation, the fixed allocation of resources to safety leads the hazard rate to explode when $\epsilon > \beta$. In a sense, $\epsilon - \beta$ characterizes the fragility of the world.

3 The (Impatient) Optimal Allocation

I now turn to the optimal allocation. I consider a representative agent that maximizes its utility. The representative agent discounts future utility with positive rate ρ: the agent is impatient. Moreover, this representative agent is selfish, i.e. it does not consider the growing population. (However, since our population is growing at a constant rate, taking into account the growing population is equivalent to lowering ρ.)

The optimal allocation of resources is a time path for c_t, h_t, s_t, ℓ_t, σ_t, A_t, B_t, M_t, δ_t that maximizes the utility of the representative agent, solving the following problem:

$$\max_{\{s_t,\ell_t,\sigma_t\}} U = \int_0^\infty M_t u(c_t) e^{-\rho t} dt, \tag{18}$$

subject to

$$c_t = A_t^\alpha \ell_t (1-\sigma_t), \tag{19}$$
$$h_t = B_t^\alpha (1-\ell_t)(1-\sigma_t), \tag{20}$$
$$\dot{A}_t = s_t^\lambda \sigma_t^\lambda N_t^\lambda A_t^\phi, \tag{21}$$
$$\dot{B}_t = (1-s_t)^\lambda \sigma_t^\lambda N_t^\lambda B_t^\phi, \tag{22}$$
$$\dot{M}_t = -\delta_t M_t, \quad \delta_t = \overline{\delta} N_t^{\epsilon-\beta} c_t^\epsilon h_t^{-\beta}. \tag{23}$$

To solve for the optimal allocation, I define the current value Hamiltonian:

$$\mathcal{H} = M_t u(c_t) + p_{at} s_t^\lambda \sigma_t^\lambda N_t^\lambda A_t^\phi + p_{bt}(1-s_t)^\lambda \sigma_t^\lambda N_t^\lambda B_t^\phi - v_t \delta_t M_t, \tag{24}$$

where s_t, ℓ_t and σ_t are our control variables and M_t, A_t, and B_t our state variables. The costate variables p_{at}, p_{bt}, and v_t capture the shadow values

of an extra consumption idea, an extra safety idea, and an extra lifetime respectively.

Based on the maximum principle and the arguments of Romer (1986), the first-order conditions characterize a solution.

It will be useful to define

$$\tilde{v}_t \equiv \frac{v_t}{u'(c_t)\, c_t}. \tag{25}$$

This is the shadow value of life, converted to consumption units by $u'(c_t)$, as a ratio to the level of consumption.

After some manipulation (see Appendix A.2) the first order conditions yield:

$$\frac{1-\ell_t}{\ell_t} = \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t}, \tag{26}$$

$$\frac{1-s_t}{s_t} = \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t} \cdot \frac{\rho - g_{p_a t} - \phi g_{At}}{\rho - g_{p_b t} - \phi g_{Bt}} \cdot \frac{g_{Bt}}{g_{At}}, \tag{27}$$

$$\frac{\sigma_t}{1-\sigma_t} = \frac{\lambda(p_{at}\dot{A} + p_{bt}\dot{B})}{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}, \tag{28}$$

$$\rho = \frac{\dot{v_t}}{v_t} + \frac{1}{v_t}[u(c_t) - v_t\delta_t], \tag{29}$$

$$\rho = \frac{\dot{p_{at}}}{p_{at}} + \frac{1}{p_{at}}[M_t u'(c_t)\alpha\frac{c_t}{A_t} + p_{at}\phi\frac{\dot{A}_t}{A_t} - \alpha\epsilon v_t M_t\frac{\delta_t}{A_t}], \tag{30}$$

$$\rho = \frac{\dot{p_{bt}}}{p_{bt}} + \frac{1}{p_{bt}}[p_{bt}\phi\frac{\dot{B}_t}{B_t} + \alpha\beta v_t M_t\frac{\delta_t}{B_t}]. \tag{31}$$

The term $\tilde{v}_t$—and in particular the product $\delta_t\tilde{v}_t$—thus determines the allocation of workers and scientists to consumption vs. safety. In Appendix A.2, I show that v_t can also be represented as

$$v_t = \frac{u(c_t)}{\rho - \delta_t + g_{vt}}, \tag{32}$$

and thus

$$\tilde{v}_t = \frac{\tilde{u}_t}{\rho - \delta_t + g_{vt}}, \quad \tilde{u}_t = \frac{u(c_t)}{u'(c_t)c_t}. \tag{33}$$

$\tilde{u}_t$ is the opportunity cost of death $u(c_t)$, converted into consumption units by $u'(c_t)$, divided by the level of consumption c_t. $\tilde{u}$ thus represents the relative

value of life. The denominator of $\tilde{v}_t$ essentially converts this into a discounted present value. Therefore, $\tilde{v}_t$ represents the discounted relative value of life and determines the demand for safety.

Note that the allocation of labor and scientists to safety is proportional to $\frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t}$. The numerator represents the marginal value of safety: the reduction in the hazard rate. The denominator represents the marginal value of consumption: the utility benefits of consumption (normalized to 1) minus the increase in the hazard rate. Note that $\delta_t\tilde{v}_t$ can't rise forever as in (as in Jones, 2016); if $\epsilon\delta_t\tilde{v}_t > 1$, the marginal value of consumption is negative.

3.1 The Optimal Allocation with $\epsilon \leq \beta$

First, consider the case in which safety goods are at least as potent in reducing existential risk as consumption goods in increasing existential risk, i.e. $\epsilon \leq \beta$. Then, existential risk weakly decreases with scale. The asymptotic growth path depends on the curvature of our preferences. The propositions here echo the results in Jones (2016).

Proposition 2. ***Optimal growth with*** $\epsilon \leq \beta$ ***and*** $\gamma > 1+(\beta-\epsilon)\left(\frac{1-\phi}{\alpha\lambda}+1\right)$ *Assume that* $\epsilon \leq \beta$ *and that the marginal utility of consumption falls rapidly, in the sense that* $\gamma > 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda}+1\right)$. *Then the optimal allocation features an asymptotic constant growth path such that as* $t \to \infty$, *the fraction of labor working in the consumption sector* ℓ_t *and the fraction of scientists working on consumption technology* s_t *both fall to zero at constant exponential*

rates, while $\sigma_t \to \sigma^*$*, and asymptotic growth is given by:*[3]

$$g_A^* = \frac{\lambda(g_s + \overline{n})}{1-\phi} > 0, \tag{34}$$

$$g_B^* = \frac{\lambda \overline{n}}{1-\phi} > g_A^*, \tag{35}$$

$$g_h^* = \overline{g}, \qquad \overline{g} \equiv \frac{\alpha\lambda\overline{n}}{1-\phi}, \tag{36}$$

$$g_c = \overline{g} \cdot \left[\frac{\beta + (\beta - \epsilon)\frac{1-\phi}{\alpha\lambda}}{\gamma + \epsilon - 1} \right] < \overline{g},\ g_c^* > 0, \tag{37}$$

$$g_\delta^* = -(\gamma - 1) g_c^* < 0, \tag{38}$$

$$g_s^* = g_\ell^* = -\overline{g} \cdot \frac{\gamma - 1 - \beta + \epsilon}{(1 + \frac{\alpha\lambda}{1-\phi})(\gamma + \epsilon - 1)} - \overline{n} \cdot \frac{\epsilon - \beta}{(1 + \frac{\alpha\lambda}{1-\phi})(\gamma + \epsilon - 1)} < 0, \tag{39}$$

$$\sigma^* = \frac{\lambda\alpha g_B}{\rho + (\gamma - 1) g_c + (1 - \phi + \lambda\alpha) g_B}. \tag{40}$$

Note that $\delta_t \to 0$ *exponentially, implying* $M_\infty > 0$*. Finally, note that this solution is valid for all* $\rho > 0$*.*

Proof. See Appendix A.3. □

Unlike in the rule of thumb allocation, the allocation of resources to safety can adjust. In particular,

$$\tilde{u}_t = \frac{u(c_t)}{u'(c_t)c_t} = \overline{u} c_t^{\gamma - 1} + \frac{1}{1-\gamma}. \tag{41}$$

Thus, given $\gamma > 1$, the relative value of life $\tilde{u}_t$ increases as consumption grows. As people grow wealthier, the marginal utility of consumption declines, and it becomes relatively more valuable to purchase more life and spend on avoiding death. Note that this happens regardless of discount rate ρ: no particular concern for the future is necessary for this dynamic. The rising value of life means that resources are shifted towards the safety sector. As such, consumption growth is substantially less than what is feasible and substantially less than safety growth.

[3]These results have the following form: $\lim_{t\to\infty} g_{ct} = g_c^*$, and so on.

Proposition 3. ***Optimal growth with*** $\epsilon < \beta$ ***and*** $\gamma < 1+(\beta-\epsilon)\left(\frac{1-\phi}{\alpha\lambda}+1\right)$ *Assume that* $\epsilon < \beta$ *and that the marginal utility of consumption falls, but not too rapidly, in the sense that* $\gamma < 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$. *Then the optimal allocation features an asymptotic constant growth path such that as* $t \to \infty$, *the fraction of labor working in the safety sector* $\tilde{\ell}_t \equiv 1-\ell_t$ *and the fraction of scientists making safety ideas* $\tilde{s}_t \equiv 1-s_t$ *both fall to 0 at constant exponential rates, while* $\sigma_t \to \sigma^*$, *and asymptotic growth is given by:*

$$g_A^* = \frac{\lambda \overline{n}}{1-\phi}, \tag{42}$$

$$g_B^* = \frac{\lambda(\overline{n} + g_{\tilde{s}})}{1-\phi} < g_A^*, g_B^* > 0, \tag{43}$$

$$g_c^* = \overline{g}, \qquad \overline{g} \equiv \frac{\alpha\lambda\overline{n}}{1-\phi}, \tag{44}$$

$$g_\delta^* = -\beta g_h^* + \epsilon g_c^* - (\beta - \epsilon)\overline{n} < 0, \tag{45}$$

with the exact values for $g_{\tilde{s}}^*$ *and* g_h^* *depending on* γ. *If* $1 < \gamma < 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$*:*

$$g_{\tilde{s}}^* = g_{\tilde{\ell}}^* = \frac{-\overline{n}\left[\frac{\alpha\lambda}{1-\phi}(1+\beta-\epsilon-\gamma) + (\beta-\epsilon)\right]}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})} < 0, \tag{46}$$

$$g_h = \overline{g} \cdot \left[1 - \frac{(1+\frac{\alpha\lambda}{1-\phi})(1-\gamma+\beta-\epsilon) + (1+\frac{1-\phi}{\alpha\lambda})(\beta-\epsilon)}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})}\right] < g_c^*. \tag{47}$$

If $\gamma \leq 1$*:*

$$g_{\tilde{s}}^* = g_{\tilde{\ell}}^* = \frac{-\overline{n}\left[(1+\frac{\alpha\lambda}{1-\phi})(\beta-\epsilon)\right]}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})} < 0, \tag{48}$$

$$g_h = \overline{g} \cdot \left[1 - \frac{(2+\frac{\alpha\lambda}{1-\phi}+\frac{1-\phi}{\alpha\lambda})(\beta-\epsilon)}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})}\right] < g_c^*. \tag{49}$$

Note that $\delta_t \to 0$ *exponentially, implying* $M_\infty > 0$.

Proof. See Appendix A.4. □

When γ is smaller, the value of life does not grow faster than the hazard rate δ_t declines. Thus, the critical product $\delta_t \tilde{v}_t$ declines, and resources are

shifted to consumption. As such, consumption growth remains as fast as is feasible.

I wish not to emphasize the difference between the allocation for larger or smaller γ, however. Instead, notice that regardless of the value of γ, the hazard rate falls exponentially to zero, and thus $M_\infty > 0$. At the same time, consumption continues to grow exponentially and $c_t \to \infty$. In that sense, the outcome of the optimal allocation in terms of the long-run future of humanity is broadly similar to the rule of thumb allocation when $\epsilon < \beta$.

To see why this is the case, note that when $\epsilon - \beta$, existential risk decreases with scale. Thus, as long as there is growth and at least some resources are allocated to safety, risk decreases.

Depending on the exact preferences, it may be possible to improve upon the rule of thumb allocation by shifting more resources to safety or to consumption over time, but the broad trajectory of the future of humanity looks bright in any case.

Finally, note that there exists a knife-edge case, which I consider for completeness.

Proposition 4. ***"Interior" growth with*** $\epsilon < \beta$ ***and*** $\gamma = 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$***, or with*** $\epsilon = \beta$ ***and*** $\gamma \leq 1$
Assume either that $\epsilon < \beta$ *and the knife-edge condition that* $\gamma = 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$*, or the knife-edge condition that* $\epsilon = \beta$ *and* $\gamma \leq 1$*. Then the optimal allocation features an asymptotic balanced growth path such that as* $t \to \infty$*,* s_t *and* ℓ_t *approach constants strictly between zero and one, and the optimal allocation features the same balanced growth path as under the rule of thumb allocation.*

Proof. See Appendix A.5. □

3.2 The Optimal Allocation with $\epsilon > \beta$

Now, consider the case where consumption goods are more potent in increasing existential risk than safety goods in reducing existential risk, i.e. $\epsilon > \beta$, but this difference is not too large, i.e. $\epsilon \not\gg \beta$. Again, the asymptotic growth path will depend on the curvature of our preferences. Here, we see a divergence from Jones (2016).

Proposition 5. ***Optimal growth with*** $\epsilon > \beta$ ***and*** $\gamma > 1$
Assume that $\epsilon > \beta$*. Assume that* $\epsilon \not\gg \beta$ *in the sense that* $\frac{\epsilon-\beta}{\beta} < \frac{\alpha\lambda}{1-\phi}$*. Finally,*

assume that the marginal utility of consumption falls rapidly, in the sense that $\gamma > 1$*. Then the optimal allocation features an asymptotic constant growth path such that as* $t \to \infty$*, the fraction of labor working in the consumption sector* ℓ_t *and the fraction of scientists working on consumption technology* s_t *both fall to zero at constant exponential rates, while* $\sigma_t \to \sigma^*$*, and asymptotic growth is given by:*

$$g_A^* = \frac{\lambda(g_s + \overline{n})}{1-\phi} > 0, \tag{50}$$

$$g_B^* = \frac{\lambda \overline{n}}{1-\phi} > g_A^*, \tag{51}$$

$$g_h^* = \overline{g}, \qquad \overline{g} \equiv \frac{\alpha\lambda\overline{n}}{1-\phi}, \tag{52}$$

$$g_c = \overline{g} \cdot \left[\frac{\beta + (\beta - \epsilon)\frac{1-\phi}{\alpha\lambda}}{\gamma + \epsilon - 1} \right] < \overline{g},\ g_c^* > 0, \tag{53}$$

$$g_\delta^* = -(\gamma - 1)g_c^* < 0, \tag{54}$$

$$g_s^* = g_\ell^* = -\overline{g} \cdot \frac{\gamma - 1 - \beta + \epsilon}{(1 + \frac{\alpha\lambda}{1-\phi})(\gamma + \epsilon - 1)} - \overline{n} \cdot \frac{\epsilon - \beta}{(1 + \frac{\alpha\lambda}{1-\phi})(\gamma + \epsilon - 1)} < 0, \tag{55}$$

$$\sigma^* = \frac{\lambda\alpha g_B}{\rho + (\gamma - 1)g_c + (1 - \phi + \lambda\alpha)g_B}. \tag{56}$$

Note that $\delta_t \to 0$ *exponentially, implying* $M_\infty > 0$*. Finally, note that this solution is valid for all* $\rho > 0$*.*

Proof. See Appendix A.6. □

Given $\gamma > 1$, the relative value of life $\tilde{u}_t$ rises as consumption grows, as before. Unlike before, however, we now have $\epsilon - \beta > 0$: existential risk grows with scale. Despite this scale effect, workers and scientists are shifted to the safety sector quickly enough that δ_t still declines exponentially on the asymptotic growth path. In turn, $M_\infty > 0$. Unlike in the rule of thumb allocation, there is a positive, nonzero probability that humanity does succumb to an existential catastrophe.

Proposition 6. ***Optimal growth with*** $\epsilon > \beta$ ***and*** $\gamma \leq 1$
Assume that $\epsilon > \beta$*. Assume that* $\epsilon \not\gg \beta$ *in the sense that* $\frac{\epsilon-\beta}{\beta} < \frac{\alpha\lambda}{1-\phi}$*. Finally, assume that the marginal utility of consumption falls, but not as rapidly, in the sense that* $\gamma \leq 1$*. Then the optimal allocation features an*

asymptotic constant growth path such that as $t \to \infty$, the fraction of labor working in the consumption sector ℓ_t and the fraction of scientists working on consumption technology s_t both fall to zero at constant exponential rates, while $\sigma_t \to \sigma^*$, and asymptotic growth is given by:

$$g_A^* = \frac{\lambda(g_s + \overline{n})}{1-\phi} > 0, \tag{57}$$

$$g_B^* = \frac{\lambda \overline{n}}{1-\phi} > g_A^*, \tag{58}$$

$$g_h^* = \overline{g}, \qquad \overline{g} \equiv \frac{\alpha \lambda \overline{n}}{1-\phi}, \tag{59}$$

$$g_c^* = \overline{g} - (\overline{n} + \overline{g})\frac{\epsilon - \beta}{\epsilon} < g_h^*,\ g_c^* > 0, \tag{60}$$

$$g_\delta^* = 0, \tag{61}$$

$$\delta_t \to \frac{(1-\gamma)\rho + (1-\gamma)^2 g_c}{\epsilon + 1 - \gamma}, \tag{62}$$

$$g_s^* = g_\ell^* = -\frac{\epsilon - \beta}{\epsilon}\overline{n} < 0. \tag{63}$$

Note that in this case $M_\infty = 0$.

Proof. See Appendix A.7. □

Unlike when $\epsilon < \beta$, in this case when $\epsilon > \beta$, workers and scientists are shifted to safety even when $\gamma \leq 1$. This is because even though the relative value of life $\tilde{u}_t$ is bounded when $\gamma \leq 1$, δ_t continues increasing because of the scale effect, so $\tilde{v}_t \delta_t$ increases. Nevertheless, despite resources being shifted to safety, they are not shifted to safety quickly enough to bound $\int_0^\infty \delta_s ds$, so the long-run probability of humanity's survival is $M_\infty = 0$ when $\gamma \leq 1$.

Note that unlike in Jones (2016), it is now optimal for scientists and workers to shift exponentially from the consumption sector to the safety sector for all γ, not just the narrower class of preferences with γ significantly greater than one.

More importantly, however, consider the comparison of the optimal allocation to the rule of thumb allocation. In the rule of thumb allocation when $\epsilon > \beta$, $\delta_t \to \infty$ and $M_\infty = 0$ because of the scale effect of existential risk. By contrast, resources are shifted to the safety sector in optimal allocation, counteracting the scale effect. Thus, δ_t converges to a small constant or even

zero. In fact, given $\gamma > 1$, the optimal allocation features δ_t falling to zero exponentially, and thus $M_\infty > 0$.

The case of $\epsilon > \beta$ is thus a world in which existential risk is an enormous challenge, but can still be overcome. With a static concern for safety, as in the rule of thumb allocation, the scale effect portends disaster. By shifting resources to safety, as in the optimal allocation, this scale effect can be contained; in fact, when $\gamma > 1$, even the impatient optimal allocation features a nonzero probability of humanity's survival in the long run.

3.3 Certain Existential Catastrophe with $\epsilon \gg \beta$

Now, consider the case in which $\epsilon \gg \beta$, i.e. consumption goods are significantly more potent in increasing existential risk than safety goods are in reducing it.

Proposition 7. ***No asymptotic growth path with*** $\epsilon \gg \beta$
Assume that $\epsilon \gg \beta$ *in the sense that* $\frac{\epsilon-\beta}{\beta} > \frac{\alpha\lambda}{1-\phi}$. *Assume reasonable preferences in the sense that there is some (arbitrarily low) level of consumption below which dying is preferring to living (i.e.* $\gamma \geq 1$, *or* $\gamma < 1$ *and* $\overline{u} < 0$*). Then there is no asymptotic growth path. This is true for any* ρ.

Proof. See Appendix A.8. □

To understand why this is the case, note that in the Jones (1995) version of the Romer (1990) model, growth in the long run is $\frac{\alpha\lambda}{1-\phi}\overline{n}$: the diminishing returns to R&D combined with the nonrivalry of ideas means that in the long run, the growth rate depends on the growth rate of population. In our model, even when (nearly) everyone works on safety, the contribution of safety growth to the growth rate of δ is $-\beta\frac{\alpha\lambda}{1-\phi}\overline{n}$.

At the same time, when $\epsilon > \beta$ in our model, existential risk increases with scale: population growth increases scale and thus contributes $(\epsilon - \beta)\overline{n}$ to the growth rate of δ. The problem arises when $(\epsilon - \beta)\overline{n} > \beta\frac{\alpha\lambda}{1-\phi}\overline{n}$: then, even when (nearly) everyone works on safety, they cannot stop the hazard rate δ from growing.

When $\frac{\epsilon-\beta}{\beta} > \frac{\alpha\lambda}{1-\phi}$, the scale effect of existential risk is larger than the scale effect of ideas. Thus, given exogenous population growth, there is no way to stop $\delta_t \to \infty$ and $M_\infty = 0$. Even halting population growth and stagnating would only provide temporary relief: without population growth, there is no growth in safety technology. δ_t remains at a constant high level, existential

catastrophe follows eventually, and $M_\infty = 0$. Letting fall c_t exponentially could contain existential risk, but eventually life would be so miserable that extinction would be preferable to continued existence. In short, when $\epsilon \gg \beta$, the optimal allocation is not much better than the rule of thumb allocation. Eventual existential catastrophe is inevitable regardless of what society does.

This case stands in stark contrast to the previous case when $\epsilon > \beta$. When the scale effect of existential risk is not too large, $c_t \to \infty$ and, given sufficiently curved preferences, $M_\infty > 0$. Even with a scale effect, existential risk could be overcome. When $\epsilon \gg \beta$, the scale effect of existential risk is larger than the scale effects of ideas. A key factor here is $\frac{\alpha\lambda}{1-\phi}$. If the returns to more research (ϕ) and more people working on research (λ) do not decrease as rapidly, $\frac{\alpha\lambda}{1-\phi}$ is higher, and so a larger scale effect of existential risk can be dealt with.

In some sense, the world of $\epsilon \gg \beta$ is the economist's version of the Fermi Paradox or the Doomsday Argument: the world is simply too fragile and R&D too hard for existential risk to be overcome.

Finally, note that there exists a knife-edge case in which consumption converges to a steady state. I consider this case for completeness.

Proposition 8. ***Optimal growth with*** $\frac{\epsilon-\beta}{\beta} = \frac{\alpha\lambda}{1-\phi}$
Assume that $\frac{\epsilon-\beta}{\beta} = \frac{\alpha\lambda}{1-\phi}$. *Assume reasonable preferences in the sense that there is some (arbitrarily low) level of consumption below which dying is preferring to living (i.e.* $\gamma \geq 1$, *or* $\gamma < 1$ *and* $\overline{u} < 0$*). Then the optimal allocation features an asymptotic growth path in which* s_t *and* ℓ_t *fall to zero exponentially,* $c_t \to c^*$, $g_\delta = 0$, *and* $M_\infty = 0$.

Proof. See Appendix A.9. □

3.4 Summary

To provide an overview of the various optimal allocations, I have compiled an overview of the asymptotic growth paths under different parameter values below. For the sake of clarity, I have omitted the knife-edge cases.

Table 1: Overview of Optimal Allocations

	$\epsilon < \beta$	$\epsilon > \beta$	$\epsilon \gg \beta$
Rule of thumb allocation	$g_c = g_h = \overline{g}$ $\delta_t \to 0$ $M_\infty > 0$	$g_c = g_h = \overline{g}$ $\delta_t \to \infty$ $M_\infty = 0$	
Optimal allocation with small γ	$s_t, \ell_t \to 1$ $g_c = \overline{g}$, $g_h < g_c$ $\delta_t \to 0$ $M_\infty > 0$	$s_t, \ell_t \to 0$ $g_c < g_h$, $g_h = \overline{g}$ $\delta_t \to \delta^* > 0$ $M_\infty = 0$	No asymptotic growth path. $M_\infty = 0$
Optimal allocation with large γ	$s_t, \ell_t \to 0$ $g_c < g_h$, $g_h = \overline{g}$ $\delta_t \to 0$ $M_\infty > 0$	$s_t, \ell_t \to 0$ $g_c < g_h$, $g_h = \overline{g}$ $\delta_t \to 0$ $M_\infty > 0$	
Existential risk in rule of thumb vs. optimal allocation	δ exponentially decays under rule of thumb. Optimal allocation changes pace of decay.	δ explodes under rule of thumb. Optimal allocation can contain growth in δ.	Doomed to existential catastrophe whatever society does.

4 Evidence on Parameters

To understand our results, we have to know what realistic parameter values are and thus which world we live in. In particular, it would be very helpful to know whether $\epsilon > \beta$.

Evidence from broad trends in growth and existential risk

Over the past century, world economic output has grown manyfold. Technological risk to human civilization has arguably grown manyfold as well. Nuclear winter, catastrophic climate change, and genetically-engineered pandemics are all risks that have emerged in the past century. The Bulletin of Atomic Scientists, who publishes the "Doomsday Clock" assessing the likelihood of existential catastrophe, puts it as follows:

> Our species has never before in its 200,000-year history been so close to a disaster as we are this century. Its unsettling enough that the Doomsday Clock has been set to an ominous 3 minutes to midnight (or doom) since 2015 [Note: 2 minutes to midnight since 2018]. But the real gravity of our situation only comes into

> focus once one realizes that before 1945, there was no need for the Doomsday Clock in the first place, given the low probability of doom. (Torres, 2016)

The key question then is what has happened to the fraction of safety spending as fraction of total output. If safety spending has not decreased as a fraction of total output, the functional form $\delta_t = \overline{\delta} C_t^{\epsilon} H_t^{\beta}$ immediately implies $\epsilon > \beta$.

Regrettably, I have not been able to find good data on safety spending as it is defined in this model. However, it seems like the effort spent on mitigating existential risk was approximately none a century or two ago. For example, nuclear disarmament and security, climate change abatement, and research on AI safety are all efforts that began only in the past century, and these efforts appear to be intensifying in the past decades. In that sense, it appears that the fraction of output spent on safety has increased.

It would clearly be desirable to collect better data on both the level of existential risk and the fraction of output spent on safety. Nevertheless, the general trend appears to imply $\epsilon > \beta$.

Evidence on $\frac{\alpha\lambda}{1-\phi}$

Jones and Romer (2010) give a broad and plausible range of $\frac{\alpha\lambda}{1-\phi} \in [1/2, 2]$. Note that this figure is for the economy as a whole. The R&D production function for the safety sector may be different, although we have no reason to believe it is, and so our base case should be that the nature of R&D is similar in both sectors. As such, I have imposed the same parameters on both the consumption and safety ideas production function in this paper.

Moreover, recent research by Bloom et al. (2017) demonstrates relatively sharply diminishing returns to research across a wide range of sectors using micro-evidence, indicating a low $\frac{\alpha\lambda}{1-\phi}$.

Evidence on γ

Most of the large empirical literature on the coefficient of relative risk aversion suggests $\gamma > 1$ is the relevant case. See e.g. Lucas (1994) on asset pricing and Chetty (2006) on labor supply. γ also traditionally equals the inverse of intertemporal substitution. The traditional evidence here suggests values well below one, implying γ well above 1; see Hall (2009) for a survey.

Evidence on ρ

Financial data reflecting consumer behavior tends to find pure time preferences in the range of 2%–5% (Pindyck, 2013). Weitzman (2007) finds data roughly consistent with a ρ of 2%. Nordhaus uses a rate of pure time preference of 1.5% in his seminal DICE climate change model (Nordhaus and Sztorc, 2013).

Implications

Our sparse evidence on the parameter values indicates that $\epsilon < \beta$ is unlikely, and the $\epsilon > \beta$ world may well the one we live in. At the same time, the $\epsilon \gg \beta$ case appears surprisingly possible. If we take the high-end estimate of $\frac{\alpha\lambda}{1-\phi} = 2$, then an ϵ of 3/2 and a β of 1/2 would mean that $\epsilon \gg \beta$. If we look at the lower-end estimate of $\frac{\alpha\lambda}{1-\phi} = 1/2$, which may be the more realistic number given recent evidence on the sharply diminishing returns to research, even e.g. $\epsilon = 3/4$ and $\beta < 1/2$, or $\epsilon = 1/2$ and $\beta < 1/3$, would suffice for $\epsilon \gg \beta$. Perhaps our efforts at mitigating existential risk do not matter much after all—not because existential risk isn't a problem, but because existential catastrophe is inevitable whatever we do.

It would clearly be beneficial to get better empirical evidence on these parameters. However, from now on, I will focus on the $\epsilon > \beta$ and $\epsilon \not\gg \beta$ case. This appears to be the empirically likely case. Moreover, if $\epsilon \gg \beta$, no intervention can change $M_\infty = 0$, so no intervention can unlock the astronomical value of the long-run future of humanity.

In addition, I will focus on the case where $\gamma > 1$. This appears to be the empirically relevant case. Moreover, when $\epsilon > \beta$ and $\gamma \leq 1$, we again get $M_\infty = 0$ regardless of any intervention.

5 Transition Dynamics

The analysis so far has shed light on the long-run behavior of growth and risk. However, we live in a world far away from this asymptotic result. To understand the relationship between growth and risk as it might apply to today, I consider the transition dynamics of the (impatient) optimal allocation.

In particular, I analyze the case where $\gamma > 1$ and $\epsilon > \beta$.

5.1 Laws of Motion in the Optimal Allocation

The transition dynamics of the optimal allocation can be studied as a system of six differential equations in six "state-like" variables: s_t, ℓ_t, σ_t δ_t, $y_t \equiv g_{At}$, and $z_t \equiv g_{Bt}$. With the addition of N_t, these variables then characterize all other variables. They each converge to constant values: $s^* = 0$, $\ell^* = 0$, $\sigma^* = \frac{\lambda\alpha g_B}{\rho+(\gamma-1)g_c+(1-\phi+\lambda\alpha)g_B}$, $\delta^* = 0$, $y^* = g_A$, and $z^* = g_B$.

Let $\hat{s}$ denote the growth rate of s, $\hat{\ell}$ denote the growth rate of ℓ, and so on.

Proposition 9. ***Laws of Motion in the Optimal Allocation***
In the optimal allocation, our "state-like" variables s_t, ℓ_t, σ_t δ_t, y_t, and z_t grow according to following laws of motion:

$$\hat{s} = \alpha z \frac{\lambda}{1-\lambda}(1-\ell)\frac{1-\sigma}{\sigma} - \alpha y \frac{\lambda}{1-\lambda}\frac{1-s}{s}\ell\frac{1-\sigma}{\sigma}, \tag{64}$$

$$\hat{\ell} = \frac{\theta_\ell(\text{Ⓐ} + \omega_\ell\theta_\sigma\text{Ⓑ})}{1-\omega_\ell\omega_\sigma\theta_\sigma}, \tag{65}$$

$$\hat{\sigma} = \theta_\sigma(\text{Ⓑ} + \omega_\sigma\hat{\ell}), \tag{66}$$

$$\hat{\delta} = (\epsilon-\beta)\left(\overline{n} - \hat{\sigma}\frac{\sigma}{1-\sigma}\right) + \alpha(\epsilon y - \beta z) + \hat{\ell}\left(\epsilon + \beta\frac{\ell}{1-\ell}\right), \tag{67}$$

$$\hat{y} = \lambda(\overline{n} + \hat{s} + \hat{\sigma}) - (1-\phi)y, \tag{68}$$

$$\hat{z} = \lambda\left(\overline{n} - \hat{s}\frac{s}{1-s} + \hat{\sigma}\right) - (1-\phi)z, \tag{69}$$

where the following definitions have been used:

$$\omega_\ell = (\gamma - 1 + \epsilon - \beta)\frac{\sigma}{1-\sigma}, \tag{70}$$

$$\omega_\sigma = -\left(\frac{\ell}{1-\ell}(1+\beta) + \epsilon\right), \tag{71}$$

$$\theta_\ell = \frac{(1-\ell)\left(1 + \frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)}{1 + \left(\gamma - 1 + \epsilon + \beta\frac{\ell}{1-\ell}\right)(1-\ell)\left(1 + \frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)}, \tag{72}$$

$$\theta_\sigma = \frac{1-\sigma}{1 + (\beta - \epsilon)\sigma - \lambda(1-\sigma)}, \tag{73}$$

$$Ⓐ = (\beta - \epsilon)\overline{n} + (1 - \gamma - \epsilon)\alpha y + \alpha\beta z - \rho - \delta + \frac{u(c_t)}{v_t}, \tag{74}$$

$$Ⓑ = (1-\lambda)\frac{s}{1-s}\hat{s} + (\lambda + \beta - \epsilon)\overline{n} + \alpha\beta z - \alpha\epsilon y + \frac{u(c_t)}{v_t} - \alpha\lambda z\frac{1-\ell}{1-s_t}\frac{1-\sigma_t}{\sigma_t}, \tag{75}$$

and

$$\frac{u(c_t)}{v_t} = \frac{\ell}{1-\ell}\beta\delta\tilde{u} + \epsilon\delta\tilde{u}, \qquad \tilde{u} = \overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma}, \tag{76}$$

$$c_t = \frac{\left(\frac{\delta}{\overline{\delta}}\left(\frac{\ell}{1-\ell}\left(\left(\frac{z}{y}\right)^{\frac{1}{1-\phi}}\left(\frac{s}{1-s}\right)^{\frac{\lambda}{1-\phi}}\right)^{\alpha}\right)^{-\beta}\right)^{\frac{1}{\epsilon-\beta}}}{N_t}. \tag{77}$$

Proof. See Appendix A.10. □

5.2 Numerical Simulation

Simulating this system of equations yields a candidate transition path for each set of parameters. These candidate transition paths feature two broad dynamics that emerge for different combinations of parameter values. The first dynamic features growth rates of A and B (and thus c and h) that start very high (with c very close to 0) and then fall to the steady state. The second dynamic features growth rates of A and B (and thus c and h) that start small and then rise over time to the steady state. In trying to understand the long-term dynamics of our civilization, the latter appears

to be the relevant case. Over the period of recorded history, consumption was initially broadly flat (but nonzero). Then, growth sped up. Thus, I will focus on the second case. Although the exact dynamics depend on the specific parameter values of course, the example below illustrates the central qualitative features of this case.

I set $\gamma = 1.5$, $\epsilon = 0.4$, $\beta = 0.3$, $\rho = 0.02$, and $\overline{n} = 1\%$. These are meant to be reasonable values for illustration; other values produce similar results. I choose the other parameter values, including ϕ, λ, $\overline{\delta}$, and $\overline{u}$, to target several stylized facts about the world. In particular, I seek to find a year t_0 with a value of life-year as a ratio to per capita consumption ($\tilde{u}$) of 4 (corresponding e.g. to per capita consumption of \$20,000 and a value of a life-year of \$80,000), in which consumption per capita grows at around 1 percent per year, around 95% of workers are in the consumption sector, and the hazard rate is approximately 0.1%. This 0.1% rate of existential catastrophe has become a relatively widely used benchmark; see Stern (2006) and Méjean et al. (2017, 2019).

This is not meant to be a formal calibration in any sense: we do not have good information about many of the parameters. This exercise is merely meant to illustrate the qualitative dynamics; the calibration helps us use a reasonable set of parameters. Critically, note that the qualitative dynamics of these results are similar for other parameter choices, such as different γ, different ϵ and β, and different ρ. I explain the details of the simulation in Appendix B.1.

Figure 1 shows the key allocation of workers and scientists to consumption along the transition path. Figure 2 shows the growth rates of consumption and safety along the transition path. Figure 3 shows the hazard rate δ along the transition path.

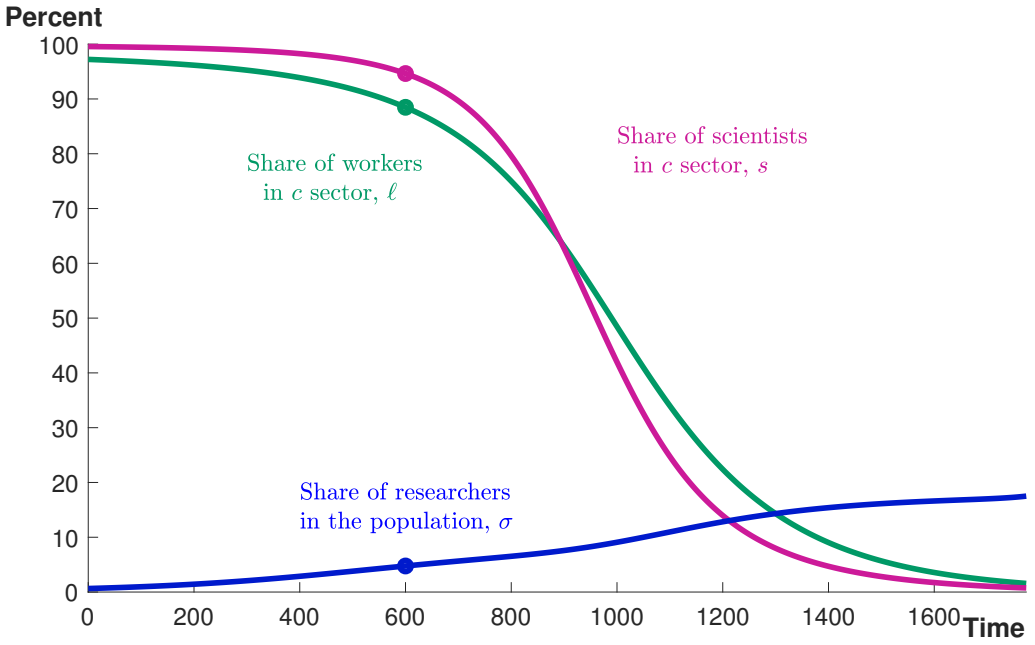

Figure 1: The allocation along the transition path. Time 600 corresponds to today and the values at this date are highlighted in the graph. A period represents a year.

Consider first the allocation variables displayed in Figure 1. At the time representing today, nearly all scientists and workers are in the consumption sector. As consumption grows and thus the relative value of life $\tilde{u}$ grows, both s and ℓ decline as resources are shifted to the safety sector. Note that initially, safety is increased by shifting workers towards the safety sector; only later are scientists shifted towards the safety sector. Both s and ℓ eventually settle in to their asymptotic, exponential decline to zero. The share of scientists in the population σ rises steadily to its steady-state value.

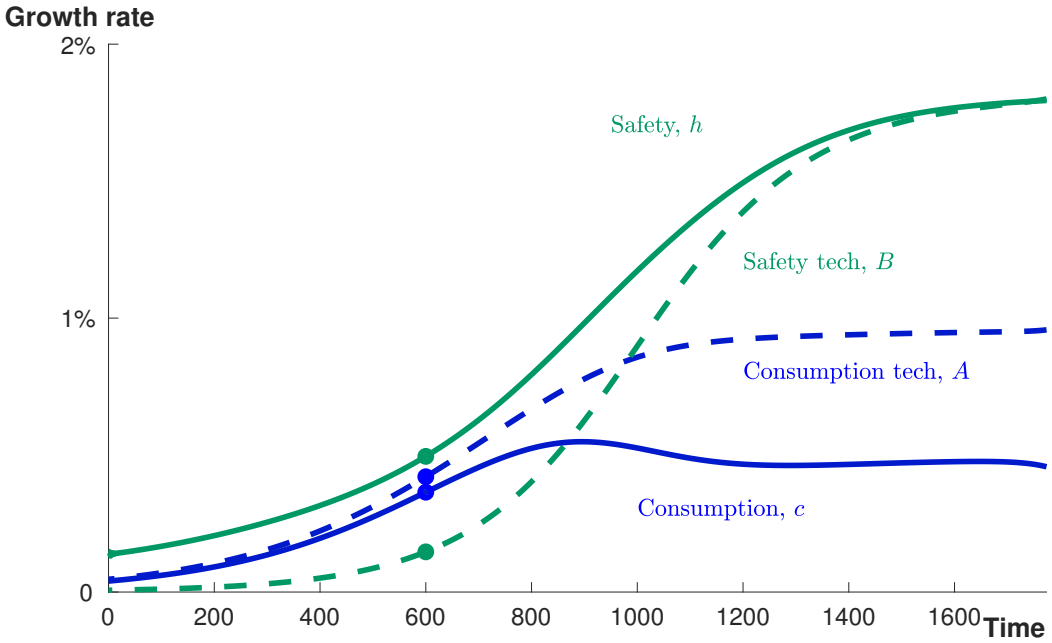

Figure 2: The growth rates along the transition path. Time 600 corresponds to today and the values at this date are highlighted in the graph. A period represents a year.

Next, consider the growth rates along the transition path in Figure 2. The growth rates of consumption technology A and thereby consumption per capita c rise steadily, accelerating from a low initial level to higher consumption growth at the time representing today. We saw that as consumption grows and the value of life rises, workers and scientists are shifted to the safety sector. This causes the growth rate of A to level off and consumption growth to slow, while the growth of safety per capita h accelerates. Note that the additional safety growth is driven by shifting workers to the safety sector at first; only after a while does the growth of safety technology B begin to accelerate. All growth rates eventually converge to their constant asymptotic values, with consumption growing significantly slower than safety. However, consumption does continue growing at a constant exponential rate.

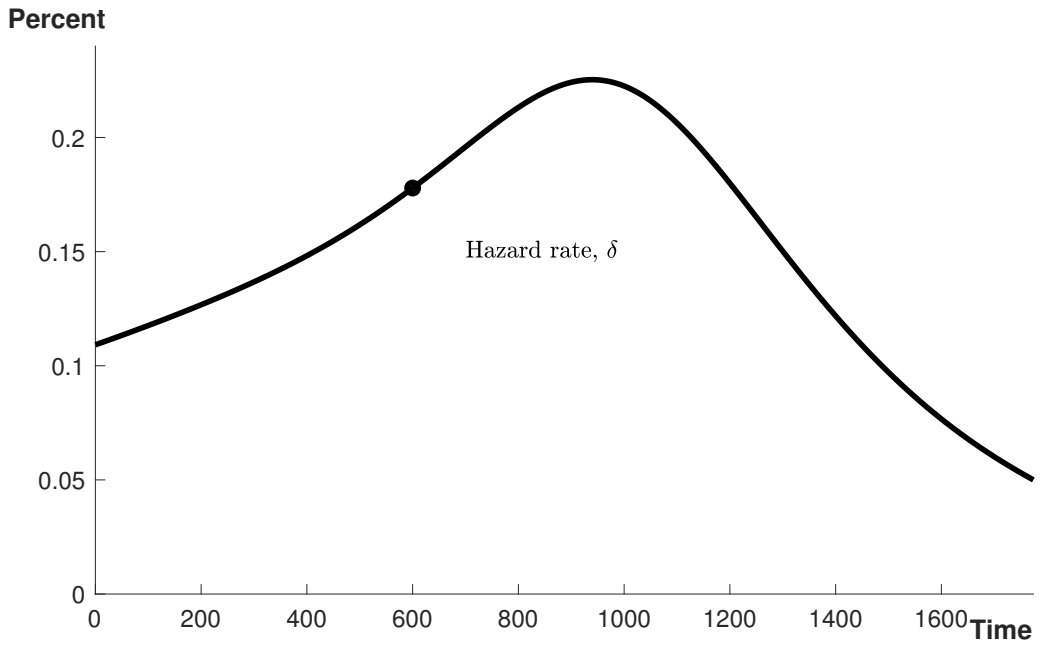

Figure 3: The hazard rate along the transition path. Time 600 corresponds to today and the value at this date is highlighted in the graph. A period represents a year.

Finally, consider perhaps the most interesting dynamic: the hazard rate along the transition path in Figure 3. The key qualitative dynamic is that the hazard rate curve has an inverted U-shape. The hazard rate starts at a relatively low level. Yet since $\epsilon > \beta$, existential risk grows with scale, so δ grows. This means that at the time representing today, the risk of an existential catastrophe is much higher than it was hundreds of years ago. As consumption grows, the value of life rises and so resources are shifted to safety. This slows the growth rate of δ, yet existential risk keeps rising—the scale effect still dominates for a while. Eventually, the growth in safety relative to consumption outpaces the scale effect, so the δ-curve bends: existential risk starts to fall. The hazard rate δ ultimately decays exponentially.

Recall that what matters in determining the long-term probability of humanity's survival is the area under the hazard rate, since $M_\infty = e^{-\int_0^\infty \delta_s ds}$. The exponential decay of the hazard rate on the asymptotic path ensures that $\int_0^\infty \delta_s ds$ is finite, and so $M_\infty > 0$. Extrapolating from the simulation, the long-run probability of human civilization's survival conditional on surviving

to the time that represents today is approximately 19.3%. However, note that I calibrated the above simulation to have a δ of approximately 0.1% today; a different calibration would rescale this curve and thus change the magnitude of the survival probability.

5.3 Discussion: The Existential Risk Kuznets Curve

This inverted U-shape of the hazard rate curve is related to the literature on the "environmental Kuznets curve," which posits an inverted U-shape relationship between economic development and pollution (see Brock and Taylor (2005) for an overview). The mechanism at work in this model is similar to the classic Stokey (1998) paper on the theory behind the environmental Kuznets curve: if $\gamma > 1$, richer societies care less about increasing consumption and more about other things, such as the environment, or, in this case, life. Initially, pollution rises with scale, but eventually declines as the relative value of environmental protection increases, producing a hump-shaped pollution curve. While the matter at hand is very different—environmental degradation that additively reduces utility versus existential catastrophes that imperil human civilization—the analogy supports the soundness of the result.

There are two important things to keep in mind, however, about what we might call the "existential risk Kuznets curve." First, the timescales involved here appear to be very long, involving hundreds or even thousands of years of economic development. Zooming in even a few hundred years around the present in the graph above, we would likely only increasing risk, much as some argue we have seen in the past century. On the one hand, this shows the value of economic theory: it allows us to gain a long-run perspective on potential societal dynamics. On the other hand, this means we cannot easily test this model prediction empirically, giving us reason for caution.

Secondly, note that this existential risk Kuznets curve appears in the transition dynamics of the *optimal* allocation. Considering that existential risk mitigation is a global public good, it is unlikely resources are allocated to safety optimally in the real world. As such, this should not be taken to be a prediction of what a particular country with a particular set of institutions will do with regard to existential risk.

Nevertheless, there are a number of reasons why we might still be interested in the transition dynamics under the (impatient) optimal allocation. For one, since there are very long timescales involved here, it is very hard to

know (and thus model) what government and societal institutions will evolve to deal with existential risk. However, the ideal these institutions will likely aim at is the optimal allocation. The optimal allocation might thus be a rough proxy for the real-world allocation.

Moreover, the (impatient) optimal allocation represents what I would call the "democratic possibilities frontier" or the "impatient public possibilities frontier." Those who are principally concerned about the long-run future of humanity and advocate for a zero rate of pure time preference might want us to spend as much as possible on safety in order to avoid existential catastrophe and enable human flourishing millions of years into the future. Indeed, even in the Hamiltonian of the optimal allocation, the relative value of life $\tilde{v}_t$ is a discounted term; the lower your discount rate ρ, the more you would want to spend on safety. However, the broader public is not so patient. As the empirical evidence cited earlier shows, people tend to have a (relatively large) positive rate of pure time preference; the public is impatient. Even perfectly designed institutions that take into account existential risk externalities will ultimately be constrained by the degree to which society actually cares about the future—they will be constrained by an impatient public. The existential risk Kuznets curve illustrates the implications of this impatience. On the one hand, this impatience results in a period of initially rising levels of risk. For example, this might mean that the arguably rising level of existential risk of the past century is not necessarily a market failure, but may well be part of the optimal path given positive pure time preference. On the other hand, rising standards of living lead even the most impatient public to start caring more about safety and averting an existential catastrophe. This leads workers and scientists to be shifted to the safety sector, eventually causing the hazard rate δ to exponentially decline. Even if people are impatient, if you make them well off enough, they will start caring about existential risk.

Seeing the arguably rising levels of existential risk in the past century, some might call for an end to economic growth. Yet this existential risk Kuznets curve indicates that stopping economic growth would be deleterious: it would simply freeze the hazard rate at a high level, leading to a fatal catastrophe sooner or later. Economic growth enables even an impatient public with a high rate of pure time preference to start caring about life, thus ultimately reducing risk and even leading to positive M_∞.

Some prominent thinkers have previously posited that humanity is passing through a unique period with an elevated risk of technological catastro-

phe. Sagan (1994) calls this the "time of perils." Parfit (2011, p. 616), concurs:

> We live during the hinge of history. Given the scientific and technological discoveries of the last two centuries, the world has never changed as fast. We shall soon have even greater powers to transform, not only our surroundings, but ourselves and our successors. If we act wisely in the next few centuries, humanity will survive its most dangerous and decisive period. Our descendants could, if necessary, go elsewhere, spreading through this galaxy.

This existential risk Kuznets curve provides theoretical evidence that grounds the intuition that we are living in a "time of perils." *We may be economically advanced enough to have created the means for our permanent destruction, but not economically advanced enough to care enough about decreasing this existential risk.*

This "time of perils" has profound implications. For instance, those alive today who care about preserving the long-term future of humanity may have extraordinary altruistic leverage. By working to reduce existential risk now (increasing the resources dedicated to safety), they can reduce the area under the "hump" of the hazard rate δ. This in turn increases M_∞, unlocking tremendous value. Moreover, since so few resources are dedicated to safety at the moment, there are likely very high marginal value opportunities available to work on safety. This is a unique situation. Suppose existential risk did not decline to zero exponentially: then $M_\infty = 0$ regardless—the existential risk curve would never bend—so reducing risk now would not change the probability of a long and flourishing future of humanity. And if existential risk did not initially increase, it would never be such a substantial challenge and there wouldn't be such high marginal value opportunities to work on reducing it.

6 Does Faster Growth Increase the Probability of Existential Catastrophe?

Faster economic growth is conventionally seen as a great boon for humanity. Yet when considering existential risk, this picture becomes more muddled.

Faster economic growth might speed up the development of potentially dangerous technology, such as powerful AI, or accelerate the pace of climate change. What if faster economic growth—in a world that does not (yet) value life highly—also accelerates the growth in risk? Could the side effect of mundane efforts to e.g. make trade more efficient or increase labor force participation be increasing the probability of an existential catastrophe? While the existential risk Kuznets curve explicated in the last section suggests we should at least want some economic growth even from the perspective of maximizing M_∞, this does not tell us anything about how the pace of economic growth affects the probability of an existential catastrophe.

First, consider a generic, uniform shock—e.g. more people working—on the balanced growth path of the rule of thumb allocation. Since the fraction of workers and scientists working in the safety sector is fixed, this increases the number of scientists and workers in the safety sector and consumption sector by the same proportion. If $\epsilon < \beta$, this shock therefore decreases the hazard rate δ. If $\epsilon = \beta$, there is no effect on δ. But if $\epsilon > \beta$, the shock increases the hazard rate δ because of the scale effect.

When $\epsilon < \beta$, faster growth reduces existential risk even in the rule of thumb allocation. Yet when $\epsilon > \beta$, accelerating growth also accelerates the growth in risk if the allocation of resources to safety does not adjust.

I will look at what happens when we accelerate growth in the (impatient) optimal allocation. In particular, I will look at the $\epsilon > \beta$ case, since in the $\epsilon < \beta$ case, faster growth reduces risk even when the allocation of resources to safety does not adjust. As explained earlier, although the real-world allocation may be imperfect, the optimal allocation might be a rough proxy for how societies will decide to allocate resources to safety in the long run. Moreover, the optimal allocation represents the "democratic possibilities frontier": the (high) positive rate of pure time preference the public appears to have dictates the degree to which societies can trade off consumption for safety. I also focus on the $\gamma > 1$ case as in the previous section.[4]

6.1 Simulating an Acceleration of Growth

First, consider what happens when growth is faster for a given time period, resulting in permanently higher economic output (i.e. this results in a per-

[4]Note that if $\gamma \leq 1$, accelerating growth would not matter for the chances of human civilization's survival in the long run: $M_\infty = 0$ anyway, regardless of whether growth is faster or slower.

manent level effect).

In this model, population growth is the driving force behind economic growth. More population means more nonrival ideas which means more output. Moreover, it easy to manipulate population growth in our model by manipulating $\overline{n}$. Thus, I consider the effect of accelerated population growth. In particular, I simulate 2% (instead of 1%) population growth for 30 years around the time representing today. We can take this to literally represent the effect of some pro-natalist policy. However, the basic dynamics illustrated below should apply to a broad class of generic accelerations in growth, e.g. increasing labor force participation, increasing human capital, increasing the number of "effective" people by making global exchange easier, or increasing research effort.

See Appendix B.2 for details of how I simulate the acceleration in growth.

The following figures compare the transition path with steady growth and the transition path with a period of accelerated growth. The transition path with steady growth is depicted with the solid colors; the transition path with a period of accelerated growth is depicted with the lighter colors. Figure 4 shows the growth rates of consumption and safety along the transition path. Figure 5 shows the fraction of workers and scientists allocated to consumption along the transition path. Figure 6 shows the relative value of life $\tilde{u}$ along the transition path. Figure 7 shows the hazard rate δ along the transition path.

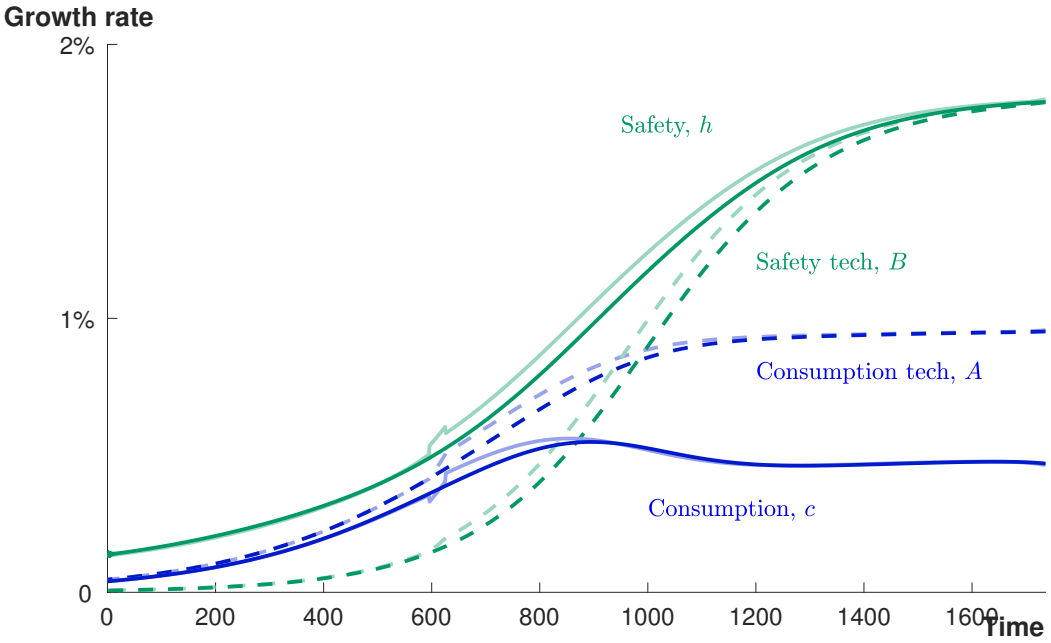

Figure 4: The growth rates along the transition path, comparing steady growth (solid colors) and a period of accelerated growth (lighter colors). A period represents a year.

Consider first the growth rates depicted in Figure 4. At around time 600, population growth accelerates for 30 years. This accelerates the growth rates of consumption technology A and safety technology B on the transition path with the accelerated growth. Both growth rates remain higher for a while, until they eventually converge to the same steady state as along the transition path with steady growth. The higher growth rates of A and B increase the growth rates of c and h. g_c and g_h are thus higher on the transition path with a period of accelerated growth, until these too converge to the same steady state as along the transition path with steady growth. Note that consumption growth actually initially decelerates a bit during the period of accelerated population growth to compensate for the scale effect of faster population growth.

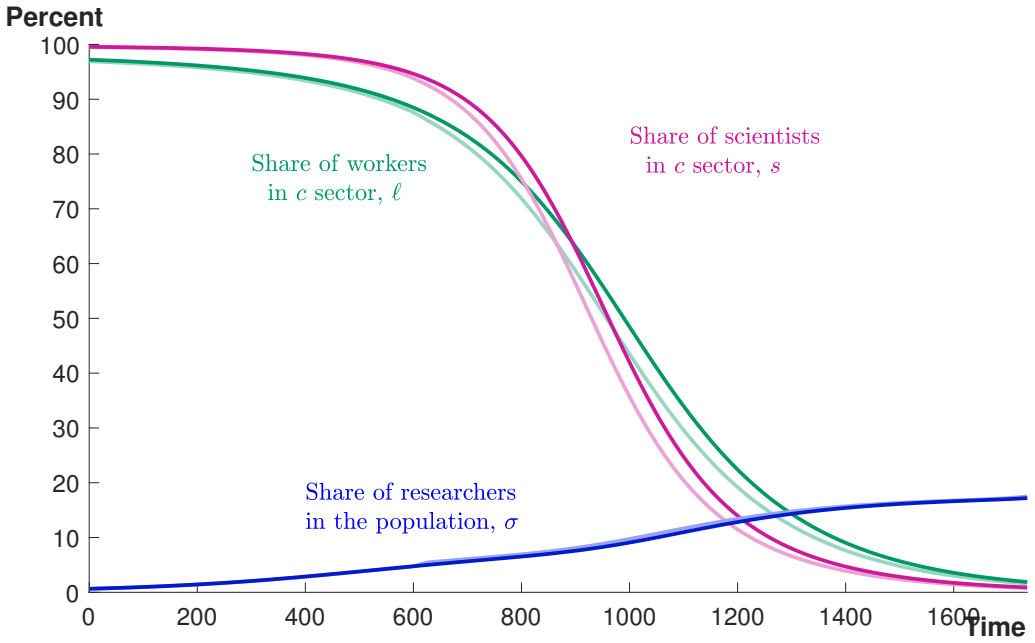

Figure 5: The allocation along the transition path, comparing steady growth (solid colors) and a period of accelerated growth (lighter colors). A period represents a year.

Next, consider the key allocation variables shown in Figure 5. The growth of the share of researchers in the population slightly increases during the period of accelerated growth. More importantly, along the transition path with the period of accelerated growth, workers and scientists are shifted to to safety earlier than along the transition path with steady growth.

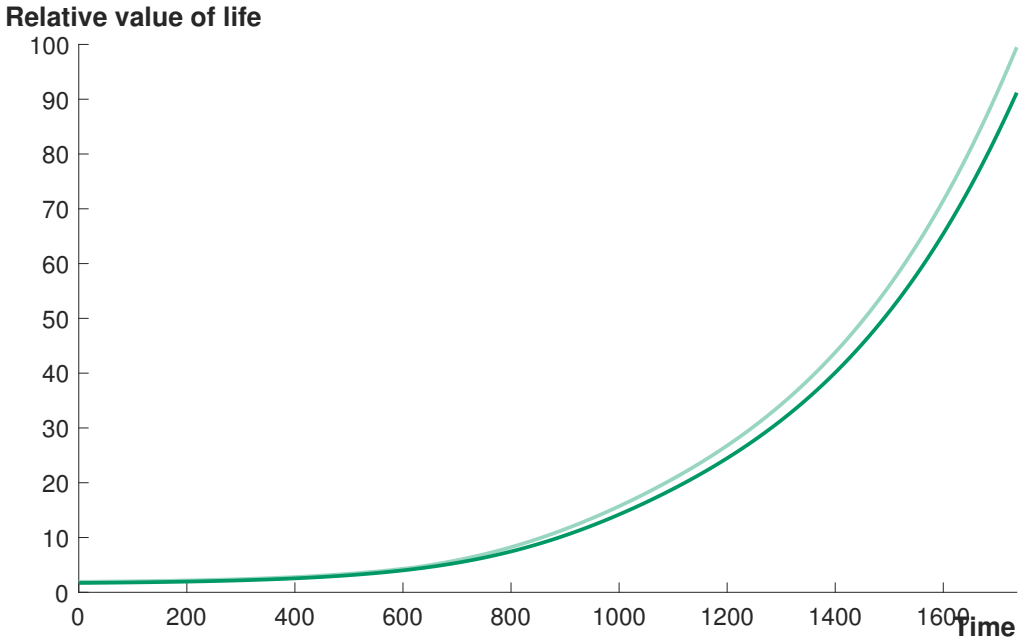

Figure 6: The relative value of life $\tilde{u}$ along the transition path, comparing steady growth (solid green) and a period of accelerated growth (light green). A period represents a year.

To understand the dynamics at play, consider Figure 6, which compares the relative value of life $\tilde{u}$ along the transition path with accelerated and steady growth. At approximately time 600, when there is a period of faster population growth, $\tilde{u}$ begins to diverge along the two transition paths. After time 600, $\tilde{u}$ is higher along the transition path with accelerated growth compared to along the transition path with steady growth. Recall the growth rates illustrated in Figure 4: the acceleration of growth meant faster consumption growth. Faster consumption growth in turn means that along the transition path with accelerated growth, people are richer, earlier, than they would have been with steady growth. Since $\tilde{u} = \frac{u(c)}{u'(c)c} = \overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma}$ and $\gamma > 1$, these richer people then value life more highly; they are more concerned for safety, earlier. Thus, resources are shifted to safety earlier, as we saw in the allocation dynamics.

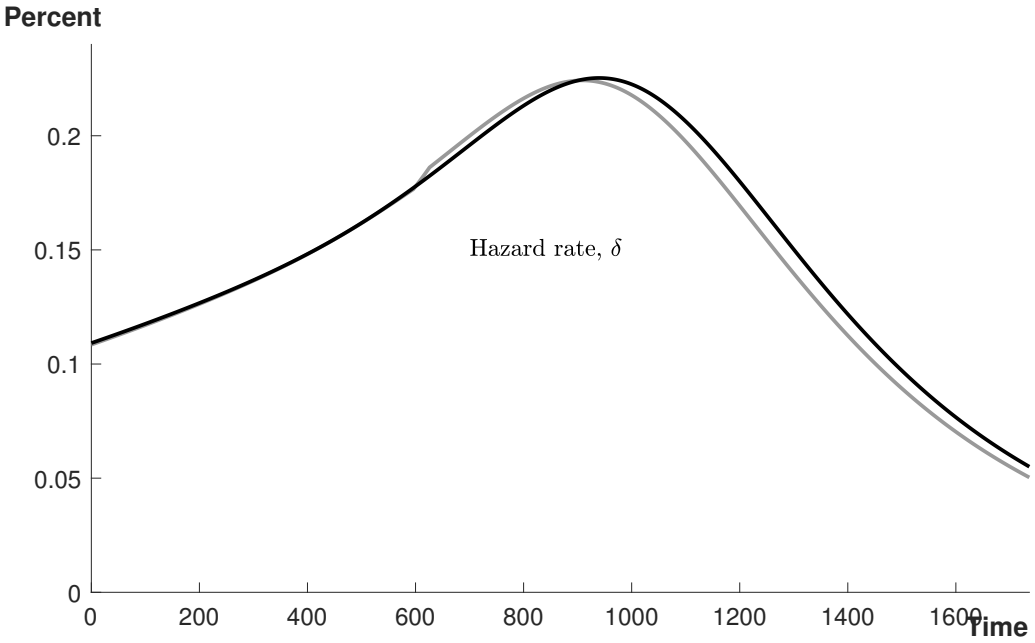

Figure 7: The hazard rate along the transition path, comparing steady growth (black) and a period of accelerated growth (gray). A period represents a year.

Consider the hazard rate δ depicted in Figure 7. After the period of accelerated growth around time 600, it initially seems as if all of the worries about faster economic growth have been confirmed: the acceleration in growth also accelerates the growth in the hazard rate. This is all an observer at the time—or even hundreds of years later—would be able to observe. Armed with empirical data, this observer would conclude that faster growth increased existential risk.

Yet zooming out, this is not so. The acceleration of growth also accelerates the rise of the relative value of life $\tilde{u}$. As such, $\tilde{v}_t$ is higher: people start caring more about safety earlier in the world with a period of accelerated growth compared to the world with steady growth. Resources are shifted to safety sooner, and thus the hazard rate curve bends earlier. In a sense, the period of faster growth accelerates the movement along the existential risk Kuznets curve. As a result, the overall area under the hazard rate curve is lower—and recall that this is all that matters for the long-run probability of civilization's survival.

As before, on the steady growth transition path, M_∞ conditional on surviving to the time that represents today is approximately 19.3%. However, on the transition path with a period of accelerated growth, M_∞ conditional on surviving to the time that represents today is approximately 20.8%. Thus, we see how the period of accelerated growth, despite increasing risk initially, improves the changes of humanity's survival in the long run! This effect is not trivial: faster growth for a relatively short period of time now appears to result in increasing the long-run probability of human survival by 1.5 percentage points. Instead of faster economic growth being a problem in the context of existential risk, this suggests that faster economic growth could actually contribute to the challenge of mitigating existential risk—even when people are impatient.

When previously discussing the existential risk Kuznets curve, I mentioned that we may well be living through a "time of perils." This analysis suggests that one way to increase the probability of humanity's survival is to simply try to get through the "time of perils" as quickly as possible. This may counterintuitively mean accelerating the increase in existential risk initially (if we are currently on the upward-sloping part of the hazard rate curve). However, this accelerationist strategy would ultimately decrease the area under the hazard rate curve and increase the probability of a long, flourishing future.

The reverse of the above happens when growth decelerates: the movement along the existential risk Kuznets curve decelerates, and society is stuck with higher levels of existential risk for longer, in turn dramatically decreasing the long-run probability of humanity's survival. Slower growth—even just for a while—doesn't just mean lower living standards, but potentially a much higher chance of an existential catastrophe and a much lower chance of a long future of humanity. This should strike fear of even short-term stagnation into the hearts of all those who care about the long-term future.

The key condition here is that $\epsilon \not\gg \beta$. The acceleration of growth initially increases risk due to the scale effect—but since the scale effect of ideas is larger than the scale effect of existential risk, it was still possible to mitigate risk eventually once $\tilde{u}$ got high enough and people started caring. Yet if $\epsilon \gg \beta$, the higher $\tilde{u}$ does not matter: even if society wanted to mitigate the additional risk later on, it would be impossible.

6.2 Simulating a Transitory Boom

So far, we have been looking at an acceleration in growth that results in a permanent level effect. What happens when we have a transitory economic boom, i.e. a time of faster growth that doesn't change the long-run level of output? For the reasons stated before, we again manipulate population growth, letting it be 2% for 40 years and then 0% for the 40 years after that (instead of a steady 1%). Thus, the long-run population is unaffected; there is simply a temporary upward blip. We may interpret this literally as the effect of a transitory baby boom. However, the basic dynamics illustrated below should apply to a broader class of transitory booms, e.g. an economic boom as part of the business cycle in which the economy is operating over capacity.

See Appendix B.2 for details of how I am simulating the acceleration in growth.

The following figures compare the transition path with steady growth and the transition path with a transitory boom. The transition path with steady growth is depicted with the solid colors; the transition path with a transitory boom is depicted with the lighter colors. Figure 8 shows the growth rates of consumption and safety along the transition path. Figure 9 shows the key allocation of workers and scientists to consumption along the transition path. Figure 10 shows the hazard rate δ along the transition path.

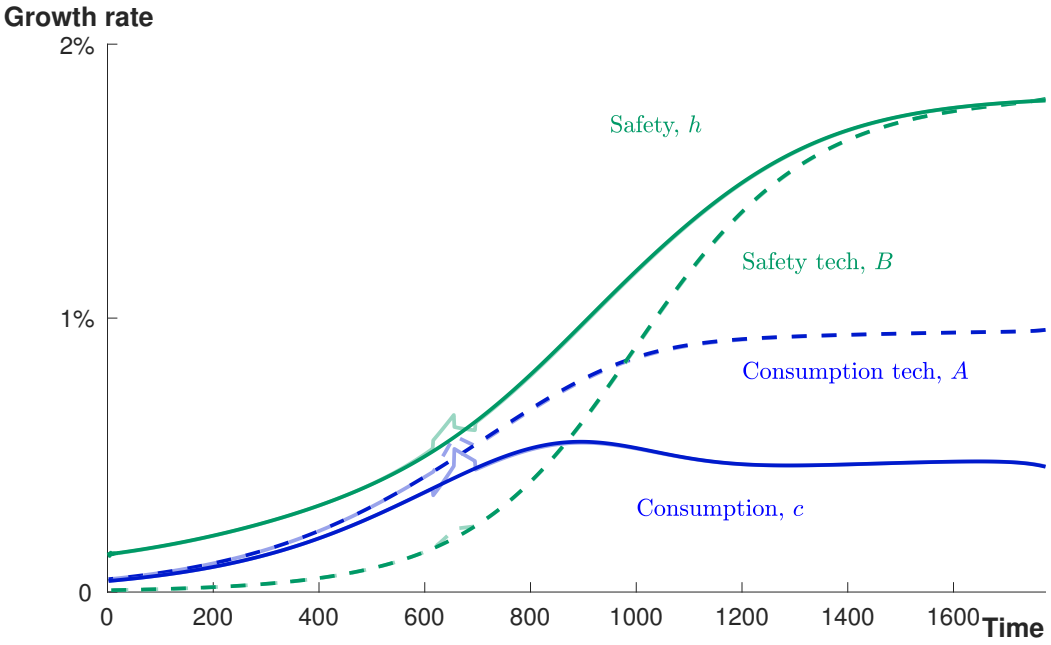

Figure 8: The growth rates along the transition path, comparing steady growth (solid colors) and a transitory boom (lighter colors). A period represents a year.

Consider first the growth rates shown in figure 8. The faster population growth initially accelerates the growth rates of both consumption technology A and safety technology B. Growth in both of these then slows down when population growth is slower during the time of slower population growth. The upward blip in the growth rates of A and B in turn lead to an upward blip in the growth rates of c and h. Nevertheless, after the temporary boom, all growth rates are the same, as had the boom not happened.

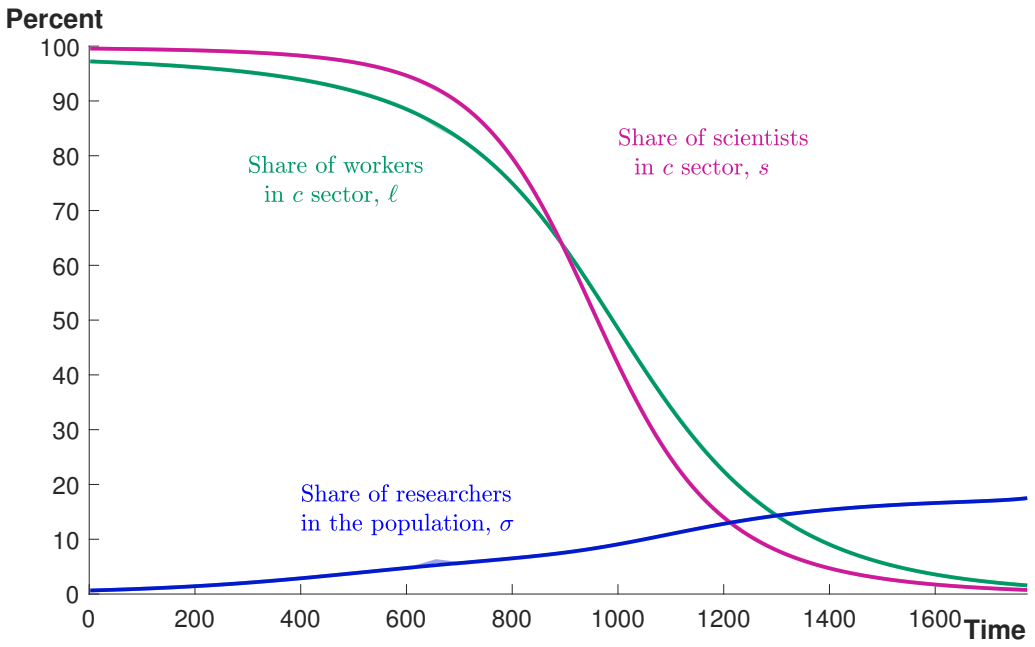

Figure 9: The allocation along the transition path, comparing steady growth (solid colors) and a transitory boom (lighter colors). A period represents a year.

Next, consider the key allocation variables depicted in figure 9. There is a temporary upward blip in the fraction of population working as researchers. There is also a temporary downward blip in the fraction of workers and researchers working in the consumption sector. Yet unlike when growth was accelerated resulting in a permanent level effect, this temporary economic boom does not change the long-term trajectory of the relative value of life; thus, the long-term trajectory of the allocation variables is unchanged.

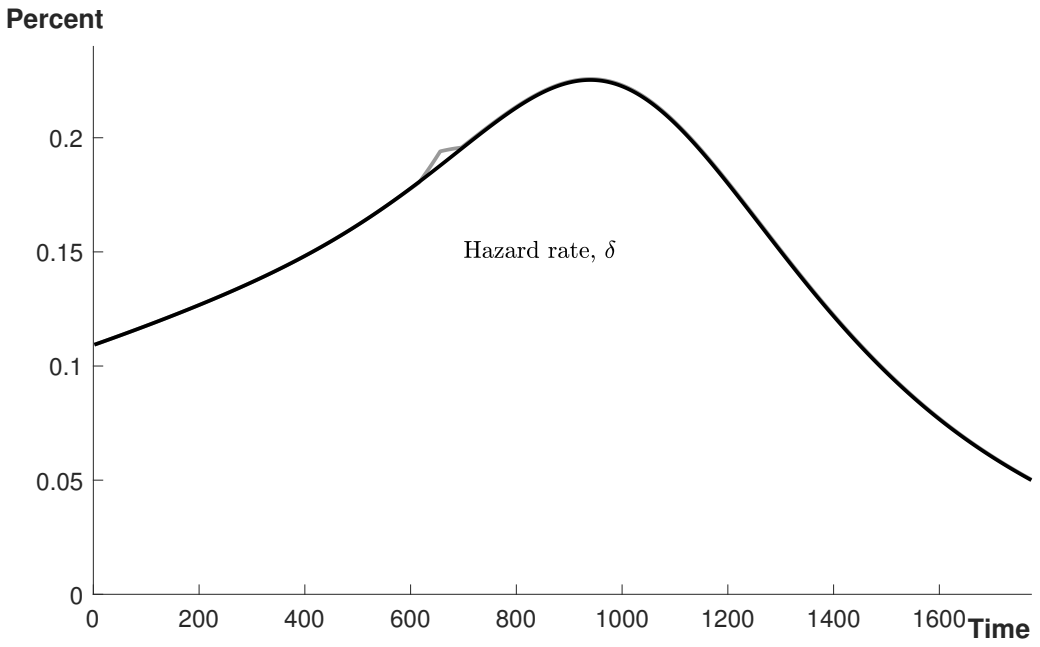

Figure 10: The hazard rate along the transition path, comparing steady growth (solid black) and a temporary boom (gray). A period represents a year.

Consider the hazard rate illustrated in figure 10. The long-run trajectory of the hazard rate is the same.

Nevertheless, this doesn't mean that temporary boom has no effect: there is an upward blip in the hazard rate during the boom. Recall again that what matters in determining the long-term probability of humanity's survival is the area under the hazard rate curve. This upward blip in the hazard curve increases the area under the hazard curve, which reduces humanity's long-term survival probability. Extrapolating from the simulation, conditional on surviving until the time representing today, the difference in long-term survival probabilities is approximately 0.17 percentage points. Considering this may just be the effect of e.g. the business cycle, and the outcome at stake is whether humanity goes extinct or not, this is again a surprisingly large effect.

The opposite occurs when we simulate a temporary bust, i.e. a slow-down in (population) growth followed by an increase in growth such that the long-term trend remains the same. Then, the hazard rate curve exhibits

a downward blip, which increases the long-term probability of humanity's survival.

We previously saw that a period of accelerating growth can increase the long-term probability of humanity's survival. Here, we thus add important nuance: the additional growth has to result in a permanent level effect. Simply "juicing" growth for a while may actually backfire, reducing the probability of humanity's survival. Nevertheless, the intuition we developed remains the same: we want to get through the "time of perils" as quickly as possible. Stagnation—in this case the "cooling off" after a transitory boom—during the "time of perils" is deleterious.

6.3 Patience vs. Growth

The key mechanism at work in this paper is that growing consumption grows people's relative value of life $\tilde{u}$ (when $\gamma > 1$). The period of accelerated growth improves the chances of civilization's survival in the long run because it accelerates the rise in the relative value of life $\tilde{u}$. As people grow richer, they care more about preventing an existential catastrophe and demand more safety.

By contrast, philosophers who are concerned about the long-term future often appeal to ethical arguments for a zero rate of pure time preference. They care about existential risk mitigation not because of a high $\tilde{u}$, but because of low or no utility discounting.

How do the these two mechanisms—increasing consumption vs. reducing the rate of pure time preference ρ—compare in terms of increasing concern for safety?

Recall that in the optimal allocation,

$$\begin{aligned}\frac{1-\ell_t}{\ell_t} &= \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t},\\ \frac{1-s_t}{s_t} &= \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t}\cdot\frac{\rho-g_{p_at}-\phi g_{At}}{\rho-g_{p_bt}-\phi g_{Bt}}\cdot\frac{g_{Bt}}{g_{At}}.\end{aligned}$$

Both the allocation of workers to safety and the allocation of scientists to safety are proportional to $\tilde{v}_t$. $\tilde{v}_t$ represents people's demand for safety. Recall that

$$\tilde{v}_t = \frac{\tilde{u}_t}{\rho-\delta_t+g_{vt}}, \quad \tilde{u}_t = \frac{u(c_t)}{u'(c_t)c_t}.$$

We see that people's concern for life depends on both $\tilde{u}$, which in turn depends on consumption, and ρ.

Thus, we can compare how lowering ρ—making people more patient—and increasing c_t—making people better off and thus increasing $\tilde{u}$—compare in terms of increasing $\tilde{v}$, people's concern for life. Although this concern for life does not necessarily translate directly to the allocation of resources to safety in the real world as it does in the optimal allocation, we would hope that the real-world allocation responds to the people's demand for safety in the long run.[5]

Proposition 10. ***Elasticities of Concern for Life***
Suppose the marginal utility of consumption falls rapidly, such that $\gamma > 1$. Let $E_\rho^{\tilde{v}}$ be the elasticity of $\tilde{v}$ with respect to ρ. Let $E_c^{\tilde{v}}$ be the elasticity of $\tilde{v}$ with respect to c_t. As $\tilde{u} \to \infty$ and $\delta \to 0$,

$$E_\rho^{\tilde{v}} \to -1, \tag{78}$$

$$E_c^{\tilde{v}} \to (\gamma - 1). \tag{79}$$

In particular, when $\tilde{u}$ is large and δ is sufficiently smaller than ρ, $E_\rho^{\tilde{v}} \approx -1$ and $E_c^{\tilde{v}} \approx (\gamma - 1)$.

Proof. See Appendix A.11. □

For large enough $\tilde{u}$ (i.e. people are already decently well off and care about life somewhat), the elasticity of $\tilde{v}$ with respect to ρ is approximately -1. Halving ρ roughly doubles the concern for life $\tilde{v}$. Moreover, the elasticity of $\tilde{v}$ with respect to c is approximately $(\gamma - 1)$. For example, if $\gamma = 2$, doubling c roughly doubles the concern for life $\tilde{v}$. The larger γ, the larger this elasticity, since a larger γ means the marginal utility of consumption decreases more rapidly and so the relative value of life $\tilde{u}$ increases more rapidly.

I have computed the approximate elasticities for different values of γ below. To help clarify the comparison, in the third column, I note what ρ would have to be reduced to, from a base of $\rho = 2\%$, to match the increase in the concern for life $\tilde{v}_t$ from a doubling of consumption.

[5]Note that when I am referring to ρ, this ρ is the rate of pure time preference without regard for increasing population. In a total utilitarian setting, the rate of pure time preference is $\rho + \overline{n}$. Thus, the elasticities with regard to the rate of pure time preference in a total utilitarian setting would be lower if $\overline{n} > 0$.

Table 2: Patience vs. Growth: Comparison of Effect on Concern for Safety

γ	$E_c^{\tilde{v}}$	ρ equivalent to doubling consumption
1.1	0.1	1.87%
1.5	0.5	1.41%
2	1	1%
4	3	0.25%

For γ close to 1, even doubling consumption is equivalent only to a small change in pure time preference in terms of regard for safety. Increasing consumption is relatively ineffective in increasing people's concern for safety. Yet for larger γ, increasing consumption has very large effect on the concern for safety—equivalent to very large reductions in pure time preference.

Note that I have been using an approximation for the elasticities that is valid for sufficiently large $\tilde{u}$. For lower $\tilde{u}$, g_{vt} is higher, so $E_\rho^{\tilde{v}}$ is lower. At the same time, for lower $\tilde{u}$, $E_c^{\tilde{v}}$, is higher—in fact, $E_c^{\tilde{v}} \to \infty$ as $\tilde{u} \to 0$. This indicates that the above numbers are a lower bound for the relative effectiveness of growing c versus lowering ρ. If people are poorer and $\tilde{u}$ is lower, increasing c is much more effective relative to decreasing ρ in increasing the concern for safety than the numbers above imply.

Nevertheless, the general takeaway is clear. Making people better off could increase concern for safety and thus demand for existential risk mitigation in a way that would be equivalent to significant changes in people's attitude toward the future.

7 Conclusion

Technological development can create or mitigate existential risks. Analyzing this in a model of endogenous growth, when the scale effect of existential risk is moderate and the marginal utility of consumption declines quickly enough, this paper grounds the intuition of some prominent thinkers that humanity may be in a critical "time of perils." We may be economically advanced enough to be able to destroy ourselves, but not economically advanced enough that we care about this existential risk and spend on safety. This "time of perils" implies that working on reducing existential risk now could be very impactful from an altruistic perspective. Faster economic growth, while initially increasing risk, can help us get through this "time of perils" more

quickly and thus increases the long-run probability of humanity's survival. Conversely, short-term economic stagnation could substantially curtail the future of human civilization. Even if you care only about the long-term future of humanity, the pace of economic growth in the short run could be key to whether we make it there. Finally, this paper also highlights the importance of the scale effect of existential risk. In particular, if this scale is larger than the scale effect of ideas, it may be impossible to avoid an existential catastrophe.

This paper suggests many future research directions. It would clearly be desirable to get better empirical data on the scale effect of existential risk. More broadly, a better understanding of how existential risk is created and mitigated would be helpful. It would also be interesting to look at the implications of a decentralized allocation, as well as possible mechanisms to efficiently provide for the global public good of existential risk mitigation. Finally, from the perspective of maximizing altruistic impact, it would be valuable to compare the impact on the long-run probability of humanity's survival from working on policies that might accelerate the rate of growth to direct work on reducing existential risk by funding the safety sector.

Appendices

A Proofs and Derivations

A.1 Proof of Proposition 1

Note that

$$g_{At} = \frac{\dot{A}_t}{A_t} = \frac{S_{at}^{\lambda}}{A_t^{1-\phi}} \tag{80}$$

Given that S_{at} is a fixed fraction of the total population, the numerator grows at rate $\lambda\overline{n}$. The denominator grows at $(1-\phi)g_{At}$. Given that on a balanced growth path, g_A must be constant, the numerator and denominator must grow at the same rate, yielding

$$g_A = \frac{\lambda\overline{n}}{1-\phi} \tag{81}$$

The same reasoning applies to g_B, giving us

$$g_B = \frac{\lambda\overline{n}}{1-\phi} \tag{82}$$

Now, note that $C_t = A^{\alpha}L_{ct}$. Given that L_{ct} is a fixed fraction of the total population, C_t grows at rate $\alpha g_{At} + \overline{n}$. Given $c_t = C_t/N_t$, c_t grows at rate αg_{At}. Thus, on the balanced growth path,

$$g_c = \alpha g_A = \frac{\alpha\lambda\overline{n}}{1-\phi} \tag{83}$$

The same reasoning applies to g_h, so

$$g_h = \alpha g_B = \frac{\alpha\lambda\overline{n}}{1-\phi} \tag{84}$$

Finally, consider what happens to $\delta_t = \overline{\delta}N_t^{\epsilon-\beta}c_t^{\epsilon}h_t^{-\beta}$. It follows directly that $g_{\delta t} = (\epsilon-\beta)\overline{n} + \epsilon g_{ct} - \beta g_{ht}$. Thus, on the balanced growth path

$$g_\delta = (\epsilon-\beta)\overline{n} + \epsilon\frac{\alpha\lambda\overline{n}}{1-\phi} - \beta\frac{\alpha\lambda\overline{n}}{1-\phi} \implies g_\delta = (\epsilon-\beta)\left(\frac{\alpha\lambda\overline{n}}{1-\phi} + \overline{n}\right) \tag{85}$$

A.2 First Order Conditions of the Hamiltonian

FOC: s_t

$$\begin{aligned}
0 &= \frac{\partial \mathcal{H}}{\partial s_t} \\
\implies 0 &= \lambda s_t^{\lambda-1} p_{at} \sigma_t^{\lambda} N_t^{\lambda} A_t^{\phi} - \lambda (1-s_t)^{\lambda-1} p_{bt} \sigma_t^{\lambda} N_t^{\lambda} B_t^{\phi} \\
\implies \lambda p_{at} \dot{A}_t s_t^{-1} &= \lambda p_{bt} \dot{B}_t (1-s_t)^{-1} \\
\implies \frac{1-s_t}{s_t} &= \frac{p_{bt}\dot{B}_t}{p_{at}\dot{A}_t}
\end{aligned} \tag{86}$$

FOC: ℓ_t

$$\begin{aligned}
0 &= \frac{\partial \mathcal{H}}{\partial \ell_t} \\
\implies 0 &= \frac{\partial}{\partial \ell_t}(M_t u(c_t) - v_t \delta_t M_t) \\
\implies M_t \frac{\partial}{\partial \ell_t}(\overline{u} + \frac{(A_t^{\alpha}\ell_t(1-\sigma_t))^{1-\gamma}}{1-\gamma}) &= M_t v_t \overline{\delta} N_t^{\epsilon-\beta} \frac{\partial}{\partial \ell_t}([A_t^{\alpha}\ell_t(1-\sigma_t)]^{\epsilon}[B_t^{\alpha}(1-\ell_t)(1-\sigma_t)]^{-\beta}) \\
\implies \frac{(1-\gamma)(A_t^{\alpha}\ell_t(1-\sigma_t))^{-\gamma}A_t^{\alpha}(1-\sigma_t)}{1-\gamma} &= v_t\overline{\delta}N_t^{\epsilon-\beta}\epsilon[A_t^{\alpha}\ell_t(1-\sigma_t)]^{\epsilon-1}A_t^{\alpha}(1-\sigma_t) \\
&\quad [B_t^{\alpha}(1-\ell_t)(1-\sigma_t)]^{-\beta} \\
&\quad + v_t\overline{\delta}N_t^{\epsilon-\beta}(-\beta)[B_t^{\alpha}(1-\ell_t)(1-\sigma_t)]^{-\beta-1} \\
&\quad B_t^{\alpha}(1-\sigma_t)(-1)[A_t^{\alpha}\ell_t(1-\sigma_t)]^{\epsilon} \\
\implies u'(c_t)c_t\ell_t^{-1} &= v_t\overline{\delta}N_t^{\epsilon-\beta}\epsilon\ell_t^{-1}c_t^{\epsilon}h_t^{-\beta} + v_t\overline{\delta}N_t^{\epsilon-\beta}\beta(1-\ell_t)^{-1}h_t^{-\beta}c_t^{\epsilon} \\
\implies u'(c_t)c_t\ell_t^{-1} &= v_t\delta_t(\epsilon\ell_t^{-1} + \beta(1-\ell_t)^{-1}) \\
\implies \frac{(1-\ell_t)}{\ell_t} &= \frac{v_t}{u'(c_t)c_t}\delta_t(\epsilon\frac{(1-\ell_t)}{\ell_t} + \beta) \\
\implies (1-\epsilon\delta_t\frac{v_t}{u'(c_t)c_t})\frac{(1-\ell_t)}{\ell_t} &= \beta\delta_t\frac{v_t}{u'(c_t)c_t} \\
\implies \frac{(1-\ell_t)}{\ell_t} &= \beta\delta_t\frac{v_t}{u'(c_t)c_t}(1-\epsilon\delta_t\frac{v_t}{u'(c_t)c_t})^{-1}
\end{aligned} \tag{87}$$

Consider the term $\frac{v_t}{u'(c_t)c_t}$: it is shadow value of life divided by the value of consumption in util terms. It thus represents the relative value of life, and it is convenient to define this explicitly:

$$\tilde{v}_t \equiv \frac{v_t}{u'(c_t)c_t} \tag{88}$$

giving us:

$$\frac{1-\ell_t}{\ell_t} = \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t} \tag{89}$$

Note that this is a very logical condition: the ratio of workers is proportional to what these workers can produce. In the numerator is the hazard rate times the relative value of life times β (the effectiveness of safety goods in reducing existential risk)—this is what can be gained by making a safety good. In the denominator is 1 (which is value of consumption relative to $\tilde{v}_t$) minus the existential risk increasing effects of consumption.

FOC: σ_t

$$\begin{aligned}
0 &= \frac{\partial \mathcal{H}}{\partial \sigma_t} \\
\implies 0 &= M_t \frac{\partial}{\partial \sigma_t}(\overline{u} + \frac{(A_t^\alpha \ell_t(1-\sigma_t))^{1-\gamma}}{1-\gamma}) + \lambda \sigma_t^{\lambda-1} p_{at} s_t^\lambda N_t^\lambda A^\phi \\
&\quad + \lambda \sigma_t^{\lambda-1} p_{at}(1-s_t)^\lambda N_t^\lambda B^\phi - M_t v_t \overline{\delta} N_t^{\epsilon-\beta} \frac{\partial}{\partial \sigma_t}([A_t^\alpha \ell_t(1-\sigma_t)]^\epsilon [B_t^\alpha (1-\ell_t)(1-\sigma_t)]^{-\beta}) \\
\implies 0 &= M_t \frac{1-\gamma}{1-\gamma}(A_t^\alpha \ell_t(1-\sigma_t))^{-\gamma} A_t^\alpha \ell_t(-1) + \frac{\lambda p_{at}\dot{A} + \lambda p_{bt}\dot{B}}{\sigma} \\
&\quad - M_t v_t \overline{\delta} N_t^{\epsilon-\beta} \epsilon [A_t^\alpha \ell_t(1-\sigma_t)]^{\epsilon-1}(-1) A_t^\alpha \ell_t [B_t^\alpha (1-\ell_t)(1-\sigma_t)]^{-\beta} \\
&\quad - M_t v_t \overline{\delta} N_t^{\epsilon-\beta}(-\beta)[B_t^\alpha (1-\ell_t)(1-\sigma_t)]^{-\beta-1}(-1) B_t^\alpha (1-\ell_t)[A_t^\alpha \ell_t(1-\sigma_t)]^\epsilon
\end{aligned}$$

$$\begin{aligned}
\implies \frac{M_t u'(c_t)c_t + v_t M_t \beta\delta_t - v_t M_t \epsilon\delta_t}{1-\sigma_t} &= \frac{\lambda(p_{at}\dot{A} + p_{bt}\dot{B})}{\sigma_t} \\
\implies \frac{\sigma_t}{1-\sigma_t} &= \frac{\lambda(p_{at}\dot{A} + p_{bt}\dot{B})}{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}
\end{aligned} \tag{90}$$

FOC: M_t

$$\rho = \frac{\partial \mathcal{H}/\partial M_t + \dot{v_t}}{v_t}$$
$$\implies \rho = \frac{\dot{v_t}}{v_t} + \frac{1}{v_t}[u(c_t) - v_t\delta_t] \tag{91}$$

FOC: A_t

$$\rho = \frac{\partial \mathcal{H}/\partial A_t + \dot{p_{at}}}{p_{at}}$$
$$\implies \rho = \frac{\dot{p_{at}}}{p_{at}} + \frac{1}{p_{at}}[M_t\frac{1-\gamma}{1-\gamma}(A_t^\alpha \ell_t(1-\sigma_t))^{-\gamma}\alpha A^{\alpha-1}\ell_t(1-\sigma_t) + \phi A^{\phi-1}p_{at}s_t^\lambda \sigma_t^\lambda N_t^\lambda$$
$$- M_t v_t \overline{\delta} N_t^{\epsilon-\beta}(B_t^\alpha(1-\ell_t)(1-\sigma_t))^{-\beta}\epsilon(A_t^\alpha \ell_t(1-\sigma_t))^{\epsilon-1}\alpha A^{\alpha-1}\ell_t(1-\sigma_t)]$$
$$\implies \rho = \frac{\dot{p_{at}}}{p_{at}} + \frac{1}{p_{at}}[M_t u'(c_t)\alpha\frac{c_t}{A_t} + p_{at}\phi\frac{\dot{A}_t}{A_t} - \alpha\epsilon v_t M_t\frac{\delta_t}{A_t}] \tag{92}$$

FOC: B_t

$$\rho = \frac{\partial \mathcal{H}/\partial B_t + \dot{p_{bt}}}{p_{bt}}$$
$$\implies \rho = \frac{\dot{p_{bt}}}{p_{bt}} + \frac{1}{p_{bt}}[\phi B^{\phi-1}p_{bt}(1-s_t)^\lambda \sigma_t^\lambda N_t^\lambda$$
$$- M_t v_t \overline{\delta} N_t^{\epsilon-\beta}(A_t^\alpha \ell_t(1-\sigma_t))^{\epsilon}(B_t^\alpha(1-\ell_t)(1-\sigma_t))^{-\beta-1}(-\beta)\alpha B^{\alpha-1}(1-\ell_t)(1-\sigma_t)]$$
$$\implies \rho = \frac{\dot{p_{bt}}}{p_{bt}} + \frac{1}{p_{bt}}[p_{bt}\phi\frac{\dot{B}_t}{B_t} + \alpha\beta v_t M_t\frac{\delta_t}{B_t}] \tag{93}$$

Transversality Conditions

Note that the three standard transversality conditions apply:

$$\lim_{t\to\infty}[e^{-\rho t}\cdot v_t M_t] = 0 \tag{94}$$
$$\lim_{t\to\infty}[e^{-\rho t}\cdot p_{at} A_t] = 0 \tag{95}$$
$$\lim_{t\to\infty}[e^{-\rho t}\cdot p_{bt} B_t] = 0 \tag{96}$$

The Price of Ideas

To solve for the allocation of scientists (see FOC: s_t), I need to solve for the relative price of ideas p_{bt}/p_{at}. To do this, I manipulate FOC B_t:

$$\begin{aligned} \rho &= \frac{\dot{p_{bt}}}{p_{bt}} + \frac{1}{p_{bt}}[p_{bt}\phi\frac{\dot{B}_t}{B_t} + \alpha\beta v_t M_t \frac{\delta_t}{B_t}] \\ \implies \rho - \frac{\dot{p_{bt}}}{p_{bt}} - \phi\frac{\dot{B}_t}{B_t} &= \frac{1}{p_{bt}}[\alpha\beta v_t M_t \frac{\delta_t}{B_t}] \\ \implies p_{bt} &= \frac{\alpha\beta v_t M_t \delta_t / B_t}{\rho - g_{p_b t} - \phi g_{Bt}} \end{aligned} \tag{97}$$

Similarly, I manipulate FOC A_t:

$$\begin{aligned} \rho &= \frac{\dot{p_{at}}}{p_{at}} + \frac{1}{p_{at}}[M_t u'(c_t)\alpha\frac{c_t}{A_t} + p_{at}\phi\frac{\dot{A}_t}{A_t} - \alpha\epsilon v_t M_t \frac{\delta_t}{A_t}] \\ \implies \rho - \frac{\dot{p_{at}}}{p_{at}} - \phi\frac{\dot{A}_t}{A_t} &= \frac{1}{p_{at}}[M_t u'(c_t)\alpha\frac{c_t}{A_t} - \alpha\epsilon v_t M_t \frac{\delta_t}{A_t}] \\ \implies p_{at} &= \frac{\alpha M_t (u'(c_t)c_t - \epsilon\delta_t v_t)/A_t}{\rho - g_{p_a t} - \phi g_{At}} \end{aligned} \tag{98}$$

Combining the two, the relative price must satisfy:

$$\frac{p_{bt}B_t}{p_{at}A_t} = \frac{\beta\delta_t v_t}{u'(c_t)c_t - \epsilon\delta_t v_t} \cdot \frac{\rho - g_{p_a t} - \phi g_{At}}{\rho - g_{p_b t} - \phi g_{Bt}} \tag{99}$$

Putting this in terms of the previously defined relative value of life $\tilde{v}_t$ I get:

$$\frac{p_{bt}B_t}{p_{at}A_t} = \frac{\beta\delta_t \tilde{v}_t}{1 - \epsilon\delta_t \tilde{v}_t} \cdot \frac{\rho - g_{p_a t} - \phi g_{At}}{\rho - g_{p_b t} - \phi g_{Bt}} \tag{100}$$

There needs to be a condition on ρ to keep the denominators positive.

Allocation of Scientists

Recall from FOC: s_t that:

$$\frac{1-s_t}{s_t} = \frac{p_{bt}\dot{B}}{p_{at}\dot{A}}$$

I can now substitute in the relative price of ideas:

$$\frac{1-s_t}{s_t} = \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t} \cdot \frac{\rho - g_{p_at} - \phi g_{At}}{\rho - g_{p_bt} - \phi g_{Bt}} \cdot \frac{g_{Bt}}{g_{At}} \tag{101}$$

Recall from FOC: ℓ_t that $(1-\ell_t)/\ell_t = (\beta\delta_t\tilde{v}_t)/(1-\epsilon\delta_t\tilde{v}_t)$, so both of these key allocation variables depend on $\delta_t\tilde{v}_t$, that is, on the race between the decline in the hazard rate and the possible rise in the value of life relative to consumption.

Note that in Jones (2016), $(1-\ell_t)/\ell_t$ and $(1-s_t)/s_t$ are instead proportional simply to $\beta\delta_t\tilde{v}_t$. Incorporating the existential risk effects of higher consumption, there is a (much) higher allocation of labor and of scientists to safety, in particular in the case that the value of life rises faster than the hazard rate falls, i.e. $\delta_t\tilde{v}_t$ rises.

Moreover, our model introduces an additional constraint. Since ℓ_t is the fraction of labor allocated to consumption, it must be that $0 < \ell_t \leq 1$ (where the strict inequality comes from the fact that at least some labor must be allocated to consumption along the balanced growth path). Thus, $\frac{(1-\ell_t)}{\ell_t}$ must be finite, i.e. the denominator cannot be 0. Given that ϵ, β, δ_t, and $\tilde{v}_t$ are guaranteed to be positive, along the optimal path:

$$\delta_t\tilde{v}_t < \frac{1}{\epsilon} \tag{102}$$

This foreshadows what will happen along the balanced growth path: given the parameters of our preferences, either δ_t falls to 0 faster than $\tilde{v}_t$, meaning $\delta_t\tilde{v}_t$ falls to 0, or $\delta_t\tilde{v}_t$ asymptotically approaches $1/\epsilon$.

Characterizing $\tilde{v}_t$

Using FOC: M_t, I obtain

$$\rho = \frac{\dot{v_t}}{v_t} + \frac{1}{v_t}[u(c_t) - v_t\delta_t]$$

$$\implies \rho - \frac{\dot{v_t}}{v_t} + \delta_t = \frac{u(c_t)}{v_t}$$

$$\implies v_t = \frac{u(c_t)}{\rho - \delta_t + g_{vt}} \tag{103}$$

$$\implies \tilde{v_t} = \frac{u(c_t)/u'(c_t)c_t}{\rho - \delta_t + g_{vt}} \tag{104}$$

Thus, the relative value of life depends on the extra utility a person enjoys versus increasing consumption on the current margin—this is why the degree of diminishing returns, γ, in our utility function plays such a key role.

Given our isoelastic CRRA utility,

$$\frac{u(c_t)}{u'(c_t)c_t} = \frac{\overline{u} + \frac{c_t^{1-\gamma}}{1-\gamma}}{c_t^{-\gamma}c_t}$$

$$\implies \frac{u(c_t)}{u'(c_t)c_t} = (\overline{u} + \frac{c_t^{1-\gamma}}{1-\gamma})(c_t^{-(1-\gamma)})$$

$$\implies \frac{u(c_t)}{u'(c_t)c_t} = \overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma} \tag{105}$$

A.3 Proof of Proposition 2

First, given equations (105) and (104) and $\gamma > 1$, along a balanced growth path in which $c_t \to \infty$:

$$\begin{aligned} g_{\tilde{v}} &= g_{\frac{u(c_t)}{u'(c_t)c_t}} - g_{\rho-\delta_t+g_{vt}} \\ g_{\tilde{v}} &= g_{\overline{u}c_t^{\gamma-1}+\frac{1}{1-\gamma}} \\ g_{\tilde{v}} &= (\gamma-1)g_c \end{aligned} \tag{106}$$

as long as δ_t converges to some constant.

I shall now conjecture that the solution for the balanced growth path takes the following form: s_t and ℓ_t fall toward zero at a constant exponential rate, while $\sigma_t \to \sigma^*$. The key condition for this result will be $\gamma > 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$.

Given $c_t = A_t^\alpha \ell_t(1-\sigma_t)$, taking logs and derivatives, in our proposed solution consumption growth is given by:

$$g_c = \alpha g_A + g_\ell \tag{107}$$

Now, observe in (101) that s_t is inversely proportional to $\frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t}$, and that the remaining terms in (101) will be constant along a balanced growth path. Observe in (FOC: ℓ_t) that ℓ_t is also inversely proportional to $\frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t}$. Thus, along the balanced growth path, $g_\ell = g_s$ and I get:

$$g_c = \alpha g_A + g_s \tag{108}$$

The growth rates of A and B follow straightforwardly from their production functions. Given $\dot{A}_t = s_t^\lambda \sigma_t^\lambda N_t^\lambda A_t^\phi$, $g_{At} = \frac{\dot{A}_t}{A_t} = \frac{s_t^\lambda \sigma_t^\lambda N_t^\lambda}{A_t^{1-\phi}}$, which becomes constant along a balanced growth path, so the numerator and denominator must grow at the same rate:

$$\begin{aligned} \lim_{t\to\infty} \dot{\ln}(s_t^\lambda \sigma_t^\lambda N_t^\lambda) &= \lim_{t\to\infty} \dot{\ln}(A_t^{1-\phi}) \\ \implies \lambda(g_s + \overline{n}) &= (1-\phi)g_A \\ \implies g_A &= \frac{\lambda(g_s+\overline{n})}{1-\phi} \end{aligned} \tag{109}$$

Given $\dot{B}_t = (1-s_t)^\lambda \sigma_t^\lambda N_t^\lambda B_t^\phi$, $g_{Bt} = \frac{\dot{B}_t}{B_t} = \frac{(1-s_t)^\lambda \sigma_t^\lambda N_t^\lambda}{B_t^{1-\phi}}$, which becomes constant a balanced growth path, so the numerator and denominator must

grow at the same rate. The key difference to A here is that s_t falling to 0 at a constant exponential rate means that $1 - s_t$ will converge to 1 and be asymptotically constant, i.e. $\lim_{t\to\infty} g_{1-s_t} = 0$.

$$\begin{aligned}\lim_{t\to\infty} \dot{\ln}((1-s_t)^\lambda \sigma_t^\lambda N_t^\lambda) &= \lim_{t\to\infty} \dot{\ln}(B_t^{1-\phi}) \\ \implies \lambda\overline{n} &= (1-\phi)g_B \\ \implies g_B &= \frac{\lambda\overline{n}}{1-\phi} \end{aligned} \tag{110}$$

Plugging (109) into (108) I thus get:

$$g_c = \alpha\frac{\lambda(g_s + \overline{n})}{1-\phi} + g_s \tag{111}$$

Plugging this into (106):

$$g_{\tilde{v}} = (\gamma - 1)[\frac{\alpha\lambda(g_s + \overline{n})}{1-\phi} + g_s] \tag{112}$$

Now make a key observation. Recall from FOC: ℓ_t that $(1-\ell_t)/\ell_t = (\beta\delta_t\tilde{v}_t)/(1-\epsilon\delta_t\tilde{v}_t)$ (and the allocation of scientists is proportional to this as well). Given a constant, positive ϵ and β, the only way for ℓ_t (and s_t) to fall to 0 is for $\delta_t\tilde{v}_t$ to grow. However, remember (102): $\delta_t\tilde{v}_t < 1/\epsilon$. Thus, as $t \to \infty$, $\delta_t\tilde{v}_t \to 1/\epsilon$, i.e. $\delta_t\tilde{v}_t$ is asymptotically constant. However, this in turn means that $\epsilon\delta_t\tilde{v}_t$ converges to 1 asymptotically, meaning that $1 - \epsilon\delta_t\tilde{v}_t$ will fall to 0 exponentially. This then delivers the desired exponential increase in $(1-\ell_t)/\ell_t$ and the exponential fall to 0 of ℓ_t (and s_t).

This convergence of $\delta_t\tilde{v}_t \to 1/\epsilon$ is unique to this model. In Jones (2016), given sufficient curvature of preferences, $\delta_t\tilde{v}_t$ goes to infinity. However, this convergence is very logical: in the denominator of our condition for $(1-\ell_t)/\ell_t$ is the marginal product of consumption labor, $1 - \epsilon\delta_t\tilde{v}_t$. 1 is the normalized value of consumption, whereas $-\epsilon\delta_t\tilde{v}_t$ is the relative impact of consumption on life. Were $\delta_t\tilde{v}_t$ to keep rising above $1/\epsilon$, the marginal product of consumption labor would be negative: consumption labor would destroy life more than it increases utility.

Thus, I know that:

$$\lim_{t\to\infty} \dot{\ln}(\delta_t \tilde{v}_t) = 0$$
$$\implies g_\delta = -g_{\tilde{v}} \tag{113}$$

Plugging in (112):

$$g_\delta = -(\gamma - 1)[\frac{\alpha\lambda(g_s + \overline{n})}{1-\phi} + g_s] \tag{114}$$

I thus need an expression for g_δ. Given $\delta_t = \overline{\delta} N_t^{\epsilon-\beta}[A_t^\alpha \ell_t(1-\sigma_t)]^\epsilon [B_t^\alpha(1-\ell_t)(1-\sigma_t)]^{-\beta}$:

$$g_\delta = \lim_{t\to\infty} \dot{ln}(N_t^{\epsilon-\beta}[A_t^\alpha \ell_t(1-\sigma_t)]^\epsilon [B_t^\alpha(1-\ell_t)(1-\sigma_t)]^{-\beta})$$
$$\implies g_\delta = \lim_{t\to\infty} \left[(\epsilon-\beta)\dot{\ln}(N_t) + \epsilon\dot{ln}([A_t^\alpha \ell_t(1-\sigma_t)]) - \beta\dot{\ln}([B_t^\alpha(1-\ell_t)(1-\sigma_t)])\right]$$
$$\implies g_\delta = (\epsilon-\beta)\overline{n} + \epsilon(\alpha g_A + g_\ell) - \beta\alpha g_B$$
$$\implies g_\delta = \alpha(\epsilon g_A - \beta g_B) + \epsilon g_\ell + (\epsilon-\beta)\overline{n}$$
$$\implies g_\delta = \alpha(\epsilon g_A - \beta g_B) + \epsilon g_s + (\epsilon-\beta)\overline{n} \tag{115}$$

where I substitute in $g_s = g_\ell$ as explained earlier.

I plug this in and solve:

$$\alpha(\epsilon\frac{\lambda(g_s+\overline{n})}{1-\phi} - \beta\frac{\lambda\overline{n}}{1-\phi}) + \epsilon g_s + (\epsilon-\beta)\overline{n} = -(\gamma-1)[\frac{\alpha\lambda(g_s+\overline{n})}{1-\phi} + g_s]$$
$$\implies g_s(\epsilon+\gamma-1+\frac{\alpha\epsilon\lambda+\alpha\lambda(\gamma-1)}{1-\phi}) = \frac{-\alpha\lambda\epsilon\overline{n}+\alpha\beta\lambda\overline{n}-(\gamma-1)\alpha\lambda\overline{n}-(1-\phi)(\epsilon-\beta)\overline{n}}{1-\phi}$$

$$\implies g_s = \frac{-\alpha\lambda\epsilon\overline{n}+\alpha\beta\lambda\overline{n}-(\gamma-1)\alpha\lambda\overline{n}-(1-\phi)(\epsilon-\beta)\overline{n}}{(1-\phi)(\epsilon+\gamma-1)+\alpha\epsilon\lambda+\alpha\lambda(\gamma-1)}$$
$$\implies g_s = \frac{\alpha\lambda\overline{n}(-\epsilon+\beta-\gamma+1)+(1-\phi)(\epsilon-\beta)\overline{n}}{(1-\phi)(\epsilon+\gamma-1)+\alpha\lambda(\epsilon+\gamma-1)}$$
$$\implies g_s = \frac{\overline{n}[\alpha\lambda(1+\beta-\epsilon-\gamma)+(1-\phi)(\beta-\epsilon)]}{(\gamma+\epsilon-1)(\alpha\lambda-\phi+1)} \tag{116}$$

g_s is negative

$$\begin{aligned}
&\iff \alpha\lambda(1+\beta-\epsilon-\gamma)+(1-\phi)(\beta-\epsilon)<0 \\
&\qquad \iff \alpha\lambda(\gamma-1-\beta+\epsilon)>(1-\phi)(\beta-\epsilon) \\
&\iff \frac{\alpha\lambda}{1-\phi}((\gamma-1)-(\beta-\epsilon))>\beta-\epsilon \\
&\qquad \iff \frac{\alpha\lambda}{1-\phi}(\gamma-1)>(\beta-\epsilon)\left(1+\frac{\alpha\lambda}{1-\phi}\right) \\
&\qquad\qquad \iff \gamma>1+(\beta-\epsilon)\left(\frac{1-\phi}{\alpha\lambda}+1\right)
\end{aligned} \tag{117}$$

Thus, $\gamma > 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$ is the key condition delivering this balanced growth path.

I can now calculate the other asymptotic growth rates as well. It will be helpful to define:

$$\overline{g} \equiv \frac{\alpha\lambda\overline{n}}{1-\phi} \tag{118}$$

The asymptotic convergence of $1 - \ell_t$ directly implies:

$$g_h = \alpha g_B = \frac{\alpha\lambda\overline{n}}{1-\phi} = \overline{g} \tag{119}$$

I can put g_s in terms of $\overline{g}$:

$$\begin{aligned}
g_s &= \frac{\alpha\lambda\overline{n}(1+\beta-\epsilon-\gamma)+\overline{n}(1-\phi)(\beta-\epsilon)}{(\gamma+\epsilon-1)(1-\phi)+(\gamma+\epsilon-1)(\alpha\lambda)} \\
\implies g_s &= -\overline{g}\cdot\frac{\gamma-1-\beta+\epsilon}{(1+\frac{\alpha\lambda}{1-\phi})(\gamma+\epsilon-1)} - \overline{n}\cdot\frac{\epsilon-\beta}{(1+\frac{\alpha\lambda}{1-\phi})(\gamma+\epsilon-1)}
\end{aligned} \tag{120}$$

I can now calculate g_c from (111):

$$
\begin{aligned}
& g_c = \alpha \frac{\lambda(g_s + \overline{n})}{1-\phi} + g_s \\
\implies & g_c = \frac{\alpha\lambda\overline{n}}{1-\phi} + (1 + \frac{\alpha\lambda}{1-\phi}) \cdot g_s \\
\implies & g_c = \overline{g} \cdot \left[1 - \frac{(\gamma - (1+\beta-\epsilon)) + \frac{1-\phi}{\alpha\lambda}(\epsilon-\beta) + \frac{\alpha\lambda}{1-\phi}(\gamma - (1+\beta-\epsilon)) + (\epsilon-\beta)}{(1+\frac{\alpha\lambda}{1-\phi})(\gamma+\epsilon-1)} \right] \\
\implies & g_c = \overline{g} \cdot \left[1 - \frac{(1+\frac{\alpha\lambda}{1-\phi})(\gamma - (1+\beta-\epsilon)) + (1+\frac{1-\phi}{\alpha\lambda})(\epsilon-\beta)}{(1+\frac{\alpha\lambda}{1-\phi})(\gamma+\epsilon-1)} \right] \\
\implies & g_c = \overline{g} \cdot \left[1 - \frac{(\gamma - (1+\beta-\epsilon)) + (\epsilon-\beta)\frac{1+\frac{1-\phi}{\alpha\lambda}}{1+\frac{\alpha\lambda}{1-\phi}}}{\gamma-1+\epsilon} \right] \\
\implies & g_c = \overline{g} \cdot \left[1 - \frac{(\gamma - (1+\beta-\epsilon)) + (\epsilon-\beta)\frac{1-\phi}{\alpha\lambda}}{\gamma-1+\epsilon} \right] \\
\implies & g_c = \overline{g} \cdot \left[\frac{\beta + (\beta-\epsilon)\frac{1-\phi}{\alpha\lambda}}{\gamma+\epsilon-1} \right] \qquad (121)
\end{aligned}
$$

Since $g_s < 0$, $g_c < \overline{g}$. $g_c > 0$ follows directly when $\beta \geq \epsilon$, as in this case.

How Low Can ρ Go?

What values of ρ are permissible for our asymptotic growth path to be valid?

In particular, the denominators of our shadow prices must be positive and the optimal allocations must satisfy the transversality conditions.

First, consider p_{bt}. Recall that

$$p_{bt} = \frac{\alpha\beta v_t M_t \delta_t / B_t}{\rho - g_{p_b t} - \phi g_{Bt}}$$

The denominator has to be positive along the balanced growth path and $g_{p_b t} \to g_{p_b}$. Then, if the denominator is positive and thus asymptotically constant along the balanced growth path:

$$g_{p_b} = \lim_{t\to\infty} \dot{ln}(M_t) + \dot{ln}(v_t\delta_t) - g_B$$

Since δ_t converges to 0, $g_M = 0$. Moreover, recall that $\delta_t v_t = \delta_t \tilde{v}_t u'(c_t) c_t$, so $g_{\delta v} = g_{\delta \tilde{v}} + g_{u'(c)c} = g_{u'(c)c} = (1-\gamma) g_c$ since $\delta_t \tilde{v}_t$ is asymptotically constant, so

$$g_{p_b} = (1-\gamma) g_c - g_B \tag{122}$$

The condition that the denominator of g_{p_b} is positive and asymptotically constant along the balanced growth path now becomes:

$$\rho > g_{p_b} + \phi g_B = (1-\gamma) g_c + (\phi - 1) g_B \tag{123}$$

Given $\gamma > 1$ and $\phi < 1$ the right hand side is negative, meaning any $\rho \geq 0$ is valid.

Recall the transversality condition for B_t:

$$\lim_{t \to \infty} [e^{-\rho t} \cdot p_{bt} \cdot B_t] = 0$$

Note that since $\gamma > 1$, $-g_{p_b} > g_B$, so the transversality condition is satisfied even for $\rho = 0$

Now, consider p_{at}. Recall that

$$p_{at} = \frac{\alpha M_t (u'(c_t) c_t - \epsilon \delta_t v_t)/A_t}{\rho - g_{p_a t} - \phi g_{At}}$$

The denominator has to be positive along the balanced growth path and $g_{p_a t} \to g_{p_a}$. Then, if the denominator is positive and thus asymptotically constant along the balanced growth path:

$$g_{p_a} = \lim_{t \to \infty} \dot{ln}(u'(c_t) c_t - \epsilon \delta_t v_t) - g_A$$

given $g_M = 0$. Again, $\delta_t v_t = \delta_t \tilde{v}_t u'(c_t) c_t$, so

$$\begin{aligned} g_{p_a} &= \lim_{t \to \infty} \dot{ln}(1 - \epsilon \delta_t \tilde{v}_t) + \dot{ln}(u'(c_t) c_t) - g_A \\ \implies g_{p_a} &= \lim_{t \to \infty} \dot{ln}(1 - \epsilon \delta_t \tilde{v}_t) + (1-\gamma) g_c - g_A \end{aligned}$$

As $\epsilon\delta_t\tilde{v}_t$ converges to $1/\epsilon$ along the balanced growth path, $(1-\epsilon\delta_t\tilde{v}_t)$ falls exponentially to zero. Indeed, note that since $1-s_t$ and $\beta\delta_t\tilde{v}_t$ are asymptotically constant, $g_s = -g_{\beta\delta\tilde{v}/(1-\epsilon\delta\tilde{v})} = g_{1-\epsilon\delta\tilde{v}}$, so

$$g_{p_a} = g_s + (1-\gamma)g_c - g_A \tag{124}$$

The condition that the denominator of g_{p_a} be positive and asymptotically constant now becomes

$$\rho > g_{p_a} + \phi g_A = g_s + (1-\gamma)g_c + (\phi-1)g_A \tag{125}$$

Again, since $g_s < 0$, $\gamma > 1$, $\phi < 1$, $g_c > 0$, and $g_A > 0$, the right hand side is negative and any $\rho \geq 0$ is valid.

Recall the transversality condition for A_t:

$$\lim_{t\to\infty}[e^{-\rho t}\cdot p_{at}\cdot A_t] = 0$$

Since $g_s < 0$ and $\gamma > 1$, I get $-g_{p_a} > g_A$, satisfying the transversality condition for A even for $\rho = 0$.

Next, recall the final transversality condition

$$\lim_{t\to\infty}[e^{-\rho t}\cdot v_t\cdot M_t] = 0$$

Since $g_M = 0$, $M_t \to M_t^* > 0$. Thus, either g_v falls exponentially to zero or $\rho > 0$ for the transversality condition to hold.

Recall (103):

$$v_t = \frac{u(c_t)}{\rho - \delta_t + g_{vt}}$$

$u(c_t) \to \overline{u}$ and $g_{vt} \to g_v$ along a balanced growth path. Thus, given a positive denominator, the denominator is asymptotically constant as $\delta_t \to 0$, implying

$$g_v = 0 \tag{126}$$

Since δ_t falls exponentially to zero and $g_v = 0$, the denominator is positive if and only if $\rho > 0$. $\rho > 0$ then also ensures the transversality condition holds.

Thus, our balanced growth path is a valid solution for any $\rho > 0$.

Note that this ρ is still not considering population growth, i.e. considers a somewhat selfish agent.

Finally, I have to find σ^*. Substituting the prices of ideas into (FOC: σ_t) I get:

$$\begin{aligned}
&\frac{\sigma_t}{1-\sigma_t} = \frac{\lambda(p_{at}\dot{A} + p_{bt}\dot{B})}{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]} \\
\implies &\frac{\sigma_t}{1-\sigma_t} = \frac{\lambda(\frac{\alpha M_t(u'(c_t)c_t - \epsilon\delta_t v_t)/A_t}{\rho - g_{p_a t} - \phi g_{At}}\dot{A} + \frac{\alpha\beta v_t M_t \delta_t / B_t}{\rho - g_{p_b t} - \phi g_{Bt}}\dot{B})}{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]} \\
\implies &\frac{\sigma_t}{1-\sigma_t} = \lambda\alpha(g_{At}\frac{\frac{u'(c_t)c_t - \epsilon\delta_t v_t}{\rho - g_{p_a t} - \phi g_{At}}}{u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t} + g_{Bt}\frac{\frac{\beta v_t \delta_t}{\rho - g_{p_b t} - \phi g_{Bt}}}{u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t})
\end{aligned}$$

Now, recall that $\delta_t \tilde{v}_t \to 1/\epsilon$, so $\delta_t v_t = \delta_t \tilde{v}_t \cdot u'(c_t)c_t \to 1/\epsilon \cdot u'(c_t)c_t$.

$$\begin{aligned}
\implies &\frac{\sigma^*}{1-\sigma^*} = \lim_{t\to\infty} \lambda\alpha(g_{At}\frac{\frac{u'(c_t)c_t - u'(c_t)c_t}{\rho - g_{p_a t} - \phi g_{At}}}{(\beta/\epsilon)u'(c_t)c_t} + g_{Bt}\frac{\frac{(\beta/\epsilon)u'(c_t)c_t}{\rho - g_{p_b t} - \phi g_{Bt}}}{(\beta/\epsilon)u'(c_t)c_t}) \\
\implies &\frac{\sigma^*}{1-\sigma^*} = \lim_{t\to\infty} \frac{\lambda\alpha g_{Bt}}{\rho - g_{p_b t} - \phi g_{Bt}} \qquad (127)
\end{aligned}$$

I can now substitute in g_{p_b} along our balanced growth path.

$$\frac{\sigma^*}{1-\sigma^*} = \frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi)g_B} \qquad (128)$$

Therefore,

$$\begin{aligned}
\sigma^*\left[1 + \frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi)g_B}\right] &= \frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi)g_B} \\
\implies \sigma^* &= \frac{\frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi)g_B}}{1 + \frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi)g_B}} \\
\implies \sigma^* &= \frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi)g_B + \lambda\alpha g_B} \\
\implies \sigma^* &= \frac{\lambda\alpha g_B}{\rho + (\gamma-1)g_c + (1-\phi+\lambda\alpha)g_B}
\end{aligned}$$

A.4 Proof of Proposition 3

I conjecture that $\tilde{s}_t \equiv 1 - s_t$ and $\tilde{\ell}_t \equiv 1 - \ell_t$ fall exponentially to zero on the asymptotic growth path, while $\sigma_t \to \sigma^*$. Then, it follows directly that

$$g_h = \alpha g_B + g_{\tilde{\ell}} \tag{129}$$

Moreover, since $(1 - \ell_t)/\ell_t = \tilde{\ell}_t/(1 - \tilde{\ell}_t)$ and $(1 - s_t)/s_t = \tilde{s}_t/(1 - \tilde{s}_t)$ are both proportional to $(\beta\delta_t\tilde{v}_t)/(1 - \epsilon\delta_t\tilde{v}_t)$ along an asymptotic growth path, implying $g_{\tilde{s}} = g_{\tilde{\ell}}$ along the asymptotic growth path (analogous to $g_s = g_\ell$ along the asymptotic growth path in the proof of proposition 2). Thus,

$$g_h = \alpha g_B + g_{\tilde{s}} \tag{130}$$

Moreover, an asymptotically constant g_B requires

$$g_B = \frac{\lambda(\overline{n} + g_{\tilde{s}})}{1 - \phi} \tag{131}$$

thus implying

$$g_h = \alpha\frac{\lambda(\overline{n} + g_{\tilde{s}})}{1 - \phi} + g_{\tilde{s}} \tag{132}$$

Since $\tilde{\ell}_t \to 0$, ℓ_t is asymptotically constant,

$$g_c = \alpha g_A \tag{133}$$

Similarly, s_t is asymptotically constant, so an asymptotically constant g_A then directly requires that

$$g_A = \frac{\lambda\overline{n}}{1 - \phi} \tag{134}$$

thus implying

$$g_c = \alpha\frac{\lambda\overline{n}}{1 - \phi} \tag{135}$$

Now, notice that for $\tilde{s}_t$ to fall to zero exponentially, $(1-s_t)/s_t = \tilde{s}_t/(1-\tilde{s}_t)$ has to fall exponentially to zero. On the asymptotic growth path $\tilde{s}_t/(1-\tilde{s}_t)$ is

proportional to $(\beta\delta_t\tilde{v}_t)/(1-\epsilon\delta_t\tilde{v}_t)$, so for $\tilde{s}_t$ to fall to zero exponentially, $\delta_t\tilde{v}_t$ has to fall to zero exponentially, meaning $1-\epsilon\delta_t\tilde{v}_t$ is asymptotically constant. Thus, $g_{(\beta\delta_t\tilde{v}_t)/(1-\epsilon\delta_t\tilde{v}_t)} = g_{\delta_t\tilde{v}_t} = g_{\tilde{s}}$. Thus,

$$g_{\tilde{s}} = g_\delta + g_{\tilde{v}} \tag{136}$$

It straightforwardly follows that

$$\begin{aligned} & g_\delta = \epsilon g_c - \beta g_h + (\epsilon-\beta)\overline{n} \\ \implies & g_\delta = \epsilon\alpha\frac{\lambda\overline{n}}{1-\phi} - \beta\alpha\frac{\lambda(\overline{n}+g_{\tilde{s}})}{1-\phi} - \beta g_{\tilde{s}} + (\epsilon-\beta)\overline{n} \end{aligned} \tag{137}$$

To get an expression for $g_{\tilde{v}}$, I must differentiate between the cases in which $\gamma \leq 1$ and $\gamma > 1$. Note that

$$\begin{aligned} g_{\tilde{v}} &= g_{\frac{u(c_t)}{u'(c_t)c_t}} - g_{\rho-\delta_t+g_{vt}} \\ g_{\tilde{v}} &= g_{\overline{u}c_t^{\gamma-1}+\frac{1}{1-\gamma}} \end{aligned}$$

as long as δ_t converges to a constant

Thus, when $\gamma > 1$, $g_{\tilde{v}} = (\gamma-1)g_c$, while when $\gamma \leq 1$, $\tilde{v}$ is asymptotically constant so $g_{\tilde{v}} = 0$. I will consider the $\gamma > 1$ case first. Then,

$$\begin{aligned} & g_{\tilde{s}} = g_\delta + g_{\tilde{v}} \\ \implies & g_{\tilde{s}} = \epsilon\alpha\frac{\lambda\overline{n}}{1-\phi} - \beta\alpha\frac{\lambda(\overline{n}+g_{\tilde{s}})}{1-\phi} - \beta g_{\tilde{s}} + (\epsilon-\beta)\overline{n} + (\gamma-1)\alpha\frac{\lambda\overline{n}}{1-\phi} \\ \implies & g_{\tilde{s}} = \frac{\alpha\lambda\overline{n}}{1-\phi}(\epsilon-\beta+\gamma-1) - \beta g_{\tilde{s}}(1+\frac{\alpha\lambda}{1-\phi}) + (\epsilon-\beta)\overline{n} \\ \implies & g_{\tilde{s}}(1+\beta+\beta\frac{\alpha\lambda}{1-\phi}) = \frac{\alpha\lambda\overline{n}}{1-\phi}(\epsilon-\beta+\gamma-1) + (\epsilon-\beta)\overline{n} \\ \implies & g_{\tilde{s}} = \frac{\overline{n}\left[\alpha\lambda(\epsilon-\beta+\gamma-1) + (1-\phi)(\epsilon-\beta)\right]}{(1+\beta)(1-\phi)+\beta\alpha\lambda} \\ \implies & g_{\tilde{s}} = \frac{-\overline{n}\left[\frac{\alpha\lambda}{1-\phi}(1-\epsilon+\beta-\gamma) + (\beta-\epsilon)\right]}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})} \end{aligned} \tag{138}$$

Then, the condition for $g_{\tilde{s}}$ to be negative is

$$\begin{aligned}
&\alpha\lambda(\epsilon-\beta+\gamma-1)+(1-\phi)(\epsilon-\beta)<0\\
\iff &\frac{\alpha\lambda}{1-\phi}((\gamma-1)-(\beta-\epsilon))<(\beta-\epsilon)\\
\iff &\frac{\alpha\lambda}{1-\phi}(\gamma-1)<(\beta-\epsilon)\left(1+\frac{\alpha\lambda}{1-\phi}\right)\\
\iff &\gamma<1+(\beta-\epsilon)\left(\frac{1-\phi}{\alpha\lambda}+1\right) \qquad (139)
\end{aligned}$$

which is the key condition delivering our result.

Let $\overline{g}\equiv\frac{\alpha\lambda\overline{n}}{1-\phi}$. I can now calculate g_h:

$$\begin{aligned}
&g_h=\frac{\alpha\lambda\overline{n}}{1-\phi}+\left(1+\frac{\alpha\lambda}{1-\phi}\right)g_{\tilde{s}}\\
\implies &g_h=\frac{\alpha\lambda\overline{n}}{1-\phi}+\left(1+\frac{\alpha\lambda}{1-\phi}\right)\frac{\overline{n}\left[\frac{\alpha\lambda}{1-\phi}(\epsilon-\beta+\gamma-1)+(\epsilon-\beta)\right]}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})}\\
\implies &g_h=\overline{g}\left(1+\frac{\frac{\alpha\lambda}{1-\phi}(\epsilon-\beta+\gamma-1)+(\epsilon-\beta)+(\epsilon-\beta+\gamma-1)+\frac{1-\phi}{\alpha\lambda}(\epsilon-\beta)}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})}\right)\\
\implies &g_h=\overline{g}\cdot\left[1-\frac{(1+\frac{\alpha\lambda}{1-\phi})(1-\gamma+\beta-\epsilon)+(1+\frac{1-\phi}{\alpha\lambda})(\beta-\epsilon)}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})}\right] \qquad (140)
\end{aligned}$$

In the case that $\gamma\leq 1$, I get

$$\begin{aligned}
&g_{\tilde{s}}=g_{\delta}+g_{\tilde{v}}\\
\implies &g_{\tilde{s}}=\epsilon\alpha\frac{\lambda\overline{n}}{1-\phi}-\beta\alpha\frac{\lambda(\overline{n}+g_{\tilde{s}})}{1-\phi}-\beta g_{\tilde{s}}+(\epsilon-\beta)\overline{n}\\
\implies &g_{\tilde{s}}=\frac{\alpha\lambda\overline{n}}{1-\phi}(\epsilon-\beta)-\beta g_{\tilde{s}}(1+\frac{\alpha\lambda}{1-\phi})+(\epsilon-\beta)\overline{n}\\
\implies &g_{\tilde{s}}(1+\beta+\beta\frac{\alpha\lambda}{1-\phi})=\frac{\alpha\lambda\overline{n}}{1-\phi}(\epsilon-\beta)+(\epsilon-\beta)\overline{n}\\
\implies &g_{\tilde{s}}=\frac{\overline{n}\left[\alpha\lambda(\epsilon-\beta)+(1-\phi)(\epsilon-\beta)\right]}{(1+\beta)(1-\phi)+\beta\alpha\lambda}\\
\implies &g_{\tilde{s}}=\frac{-\overline{n}\left[(1+\frac{\alpha\lambda}{1-\phi})(\beta-\epsilon)\right]}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})} \qquad (141)
\end{aligned}$$

which given $\beta > \epsilon$ is negative as conjectured.

I can then calculate g_h:

$$
\begin{aligned}
& g_h = \frac{\alpha\lambda\overline{n}}{1-\phi} + \left(1 + \frac{\alpha\lambda}{1-\phi}\right) g_{\tilde{s}} \\
\implies & g_h = \frac{\alpha\lambda\overline{n}}{1-\phi} + \left(1 + \frac{\alpha\lambda}{1-\phi}\right) \frac{\overline{n}\left[(1+\frac{\alpha\lambda}{1-\phi})(\epsilon-\beta)\right]}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})} \\
\implies & g_h = \overline{g} \cdot \left[1 - \frac{(2+\frac{\alpha\lambda}{1-\phi}+\frac{1-\phi}{\alpha\lambda})(\beta-\epsilon)}{1+\beta(1+\frac{\alpha\lambda}{1-\phi})}\right]
\end{aligned}
$$

Finally, note that given $g_{\tilde{v}} \geq 0$ and $g_{\tilde{s}} < 0$ in both cases, since $g_{\tilde{v}} + g_\delta = g_{\tilde{s}}$, I know g_δ is negative, and thus I know that δ_t falls exponentially to zero.

A.5 Proof of Proposition 4

In the case that $\epsilon < \beta$, the proof is straightforward. In particular, in the previous two proofs for the cases that $\gamma > 1$, when I plug in $\gamma = 1 + (\beta - \epsilon)\left(\frac{1-\phi}{\alpha\lambda} + 1\right)$ it immediately follows that $g_s = 0$.

In the case that $\epsilon = \beta$ and $\gamma \leq 1$, the proof is straightforward as well. In particular, consider the $\gamma \leq 1$ case in the previous proof; plugging in $\epsilon = \beta$ immediately yields $g_s = 0$.

Once $g_s = 0$, the proof proceeds as in the rule of thumb allocation.

A.6 Proof of Proposition 5

This proof is essentially the same as for proposition 2, including the section on the minimum valid value of ρ, with slight modifications.

First, $g_s < 0$ follows directly from $\gamma > 1$ and $\epsilon > \beta$, so no additional condition is necessary.

Second, I need to ensure that $g_c > 0$, which is necessary since I am assuming $c_t \to \infty$ such that $g_{\bar{v}} = 0$. If $\epsilon \gg \beta$, this is not the case. Specifically,

$$\begin{aligned} g_c = \overline{g} \cdot \left[\frac{\beta + (\beta - \epsilon)\frac{1-\phi}{\alpha\lambda}}{\gamma + \epsilon - 1} \right] > 0 \\ \iff \beta + (\beta - \epsilon)\frac{1-\phi}{\alpha\lambda} > 0 \\ \iff \beta\frac{\alpha\lambda}{1-\phi} > \epsilon - \beta \\ \iff \frac{\epsilon - \beta}{\beta} < \frac{\alpha\lambda}{1-\phi} \end{aligned} \tag{142}$$

A.7 Proof of Proposition 6

Given $\gamma \leq 1$, $\frac{u(c_t)}{u'(c_t)c_t}$ is asymptotically constant, so

$$\begin{aligned} g_{\tilde{v}} &= g_{\frac{u(c_t)}{u'(c_t)c_t}} - g_{\rho-\delta_t+g_{vt}} \\ \implies g_{\tilde{v}} &= 0 \end{aligned} \tag{143}$$

as long as δ_t converges to a constant.

Given that $\epsilon > \beta$, only $s_t \to 0$ and $\ell_t \to 0$ ensures an asymptotic growth path. $g_s = 0$ would imply $g_A = g_B$, and so $g_\delta = \alpha(\epsilon g_A - \beta g_B) + (\epsilon - \beta)\overline{n} > 0$, meaning $\delta_t \tilde{v}_t \to \infty$, which is not permissible. If $(1 - s_t) \to 0$, $g_A > g_B$, and so $g_\delta = \alpha(\epsilon g_A - \beta g_B) - \beta g_{1-s} + (\epsilon - \beta)\overline{n} > 0$, again meaning $\delta_t \tilde{v}_t \to \infty$, which is not permissible.

Thus, s_t (and ℓ_t) must fall exponentially to 0, which happens only if $\delta_t \tilde{v}_t$ rises and eventually converges asymptotically to $1/\epsilon$. Therefore, on the balanced growth path, $g_{\tilde{v}} = -g_\delta$. Thus, $0 = -g_\delta$.

The growth rates $g_\delta = \alpha(\epsilon g_A - \beta g_B) + \epsilon g_s + (\epsilon - \beta)\overline{n}$, $g_A = \frac{\lambda(g_s+\overline{n})}{1-\phi}$, and $g_B = \frac{\lambda \overline{n}}{1-\phi}$ follow just as in the proof of proposition 2.

Thus,

$$\begin{aligned} 0 = g_\delta &= \alpha(\epsilon g_A - \beta g_B) + \epsilon g_s + (\epsilon - \beta)\overline{n} \\ \implies 0 &= \alpha(\epsilon \frac{\lambda(g_s + \overline{n})}{1-\phi} - \beta \frac{\alpha\lambda\overline{n}}{1-\phi}) + \epsilon g_s + (\epsilon - \beta)\overline{n} \\ \implies g_s \cdot (-\epsilon)(1 + \frac{\alpha\lambda}{1-\phi}) &= (\epsilon - \beta)(1 + \frac{\alpha\lambda}{1-\phi})\overline{n} \\ \implies g_s &= -\frac{\epsilon - \beta}{\epsilon}\overline{n} \end{aligned} \tag{144}$$

which is negative, as conjectured.

Just like in the proof of proposition 2, $g_h = \alpha g_B = \frac{\alpha\lambda\overline{n}}{1-\phi} \equiv g$ follows.

I can then calculate g_c:

$$\begin{aligned} g_c &= \alpha g_A + g_s \\ \implies g_c &= \alpha \frac{\lambda(g_s + \overline{n})}{1-\phi} + g_s \\ \implies g_c &= \overline{n} \cdot \left[\frac{\alpha\lambda}{1-\phi} - (1 + \frac{\alpha\lambda}{1-\phi}) \frac{\epsilon - \beta}{\epsilon} \right] \\ \implies g_c &= \overline{g} - (\overline{n} + \overline{g}) \frac{\epsilon - \beta}{\epsilon} \end{aligned} \tag{145}$$

Given $\epsilon > \beta$, it follows that $g_c < \overline{g}$.

I have to check that $g_c > 0$ (since $g_{\tilde{v}} = 0$ requires $c_t \to \infty$).

$$\begin{aligned} g_c = \overline{g} - (\overline{n} + \overline{g})\frac{\epsilon - \beta}{\epsilon} &> 0 \\ \iff \epsilon\frac{\alpha\lambda}{1-\phi} &> (1 + \frac{\alpha\lambda}{1-\phi})(\epsilon - \beta) \\ \iff 0 &> (\epsilon - \beta) - \beta\frac{\alpha\lambda}{1-\phi} \\ \iff \frac{\epsilon - \beta}{\beta} &< \frac{\alpha\lambda}{1-\phi} \end{aligned} \tag{146}$$

which is the same condition that $\epsilon \not\gg \beta$ from before.

Finally, I determine $\delta_t \to \delta^*$. In the case that $\gamma < 1$:

$$\begin{aligned} \delta^* &= \frac{1}{\epsilon}\lim_{t\to\infty} 1/\tilde{v}_t \\ \implies \delta^* &= \frac{1}{\epsilon}\lim_{t\to\infty} \frac{\rho - \delta_t + g_{vt}}{\overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma}} \\ \implies \delta^* &= \frac{(\rho - \delta^* + g_v)(1-\gamma)}{\epsilon} \end{aligned}$$

Note that $v_t = \frac{u(c_t)}{\rho - \delta_t + g_{vt}}$, so given a positive and thus asymptotically constant denominator, $g_v = \lim_{t\to\infty} \frac{u'(c_t)\dot{c}_t}{u(c_t)} = \lim_{t\to\infty} \frac{u'(c_t)c}{u(c_t)} \cdot g_c = \lim_{t\to\infty} 1/(\overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma}) \cdot g_c = (1-\gamma)g_c$.

$$\begin{aligned} \implies \delta^* &= \frac{(\rho - \delta^* + (1-\gamma)g_c)(1-\gamma)}{\epsilon} \\ \implies \delta^*(1 + \frac{1-\gamma}{\epsilon}) &= \frac{(1-\gamma)\rho + (1-\gamma)^2 g_c}{\epsilon} \\ \implies \delta^* &= \frac{(1-\gamma)\rho + (1-\gamma)^2 g_c}{\epsilon(1 + \frac{1-\gamma}{\epsilon})} \\ \implies \delta^* &= \frac{(1-\gamma)\rho + (1-\gamma)^2 g_c}{\epsilon + 1 - \gamma} \end{aligned} \tag{147}$$

When $\gamma = 1$, $\frac{u(c_t)}{u'(c_t)c_t} = \overline{u} + \ln(c_t)$, so $\delta_t \to 0$. However, δ_t does not fall fast enough (not exponentially, $g_\delta = 0$)—in particular, $\delta_t \to 0$ proportional to $1/\tilde{v}_t$ proportional to $1/\ln(c_t)$—so $M_t \to 0$ even when $\gamma = 1$.

A.8 Proof of Proposition 7

First, note that I assume dying is preferred to living with no consumption, which means $\gamma \geq 1$, or $\gamma < 1$ with negative $\overline{u}$. Then if $g_c < 0$, consumption falls arbitrarily close to zero, which for our preferences then means that utility is negative, which means dying would be preferred to living with this low level of consumption. Thus, there does not exist an optimal asymptotic growth path in which $g_c < 0$.

Next, note that $\delta_t \to \infty$ cannot occur when $g_c \geq 0$ and thus $g_{\tilde{v}} \geq 0$, since then then $\delta_t \tilde{v}_t \to \infty$, which is not valid (recall that $\delta_t \tilde{v}_t$ must be less than $1/\epsilon$). Observe also that g_δ is minimized on a potential asymptotic growth path when $s_t \to 0$ asymptotically (i.e. when in the limit, everyone is working on safety).

If $\frac{\epsilon-\beta}{\beta} > \frac{\alpha\lambda}{1-\phi}$, then $(\epsilon-\beta)\overline{n} > \beta\frac{\alpha\lambda\overline{n}}{1-\phi}$. The RHS equals $\alpha\beta g_B$ when $s_t \to 0$. Given that $g_\delta = \epsilon g_c - \alpha\beta g_B + (\epsilon-\beta)\overline{n}$ on an asymptotic growth path where $s_t \to 0$, for any $g_c \geq 0$, $g_\delta > 0$, meaning that $\delta_t \to \infty$—which as just noted is not possible on an asymptotic growth path. Thus, there does not exist an asymptotic growth path.

Or put another way: if $\epsilon \gg \beta$, the scale effects pf existential risk are so large that, given exogenous population growth, even if (asymptotically) all humans worked on improving safety, the hazard rate δ would go to ∞.

A.9 Proof of Proposition 8

I conjecture that the optimal allocation features an asymptotic growth path in which $g_c = 0$ and s_t and ℓ_t decline exponentially to zero.

Then,

$$\begin{aligned} 0 = g_c = \alpha \frac{\lambda\left(g_s + \overline{n}\right)}{1-\phi} + g_s \\ \implies (1 + \frac{\alpha\lambda}{1-\phi}) g_s = -\frac{\alpha\lambda}{1-\phi}\overline{n} \\ \implies g_s = -\frac{\alpha\lambda}{(1-\phi)(1+\frac{\alpha\lambda}{1-\phi})}\overline{n} \end{aligned} \tag{148}$$

which is negative as conjectured.

Then,

$$\begin{aligned} g_\delta = \epsilon g_c + \alpha\beta g_B + (\epsilon - \beta)\overline{n} \\ \implies g_\delta = \alpha\beta \frac{\lambda\overline{n}}{1-\phi} + (\epsilon - \beta)\overline{n} \implies g_\delta = 0 \end{aligned}$$

Since $g_{\tilde{v}} = 0$ as $c_t \to c^*$, $g_\delta = 0$ then ensures that $\delta_t \tilde{v}_t$ is asymptotically constant and $\delta_t \tilde{v}_t \to 1/\epsilon$, giving us s_t falling to zero exponentially as conjectured.

A.10 Proof of Proposition 9

Note that for any variable a, $\widehat{(1-a)} = \frac{\dot{1-a}}{1-a} = -\frac{\dot{a}}{a}\frac{a}{1-a} = -\hat{a}\frac{a}{1-a}$.

Law of Motion: y

Recall that $y \equiv g_{At} = \frac{(s_t\sigma_t N_t)^\lambda}{A_t^{1-\phi}}$. Taking logs and derivatives:

$$\hat{y} = \lambda(\overline{n} + \hat{\sigma} + \hat{s}) - (1-\phi)y \tag{149}$$

Law of Motion: z

Similarly, I consider $\hat{z}$. Recall that $z \equiv g_{Bt} = \frac{((1-s_t)\sigma_t N_t)^\lambda}{B_t^{1-\phi}}$. Taking logs and derivatives:

$$\hat{z} = \lambda\left(\overline{n} + \hat{\sigma} - \hat{s}\frac{s}{1-s}\right) - (1-\phi)z \tag{150}$$

Law of Motion: δ

Recall that $\delta_t = \overline{\delta} N_t^{\epsilon-\beta} c_t^\epsilon h_t^{-\beta}$, with $c_t = A_t^\alpha \ell_t\,(1-\sigma_t)$ and $B_t^\alpha\,(1-\ell_t)\,(1-\sigma_t)$. Again, taking logs and derivatives:

$$\begin{aligned}
&\hat{\delta} = (\epsilon-\beta)\overline{n} + \epsilon g_{ct} - \beta g_{ht} \\
\implies &\hat{\delta} = (\epsilon-\beta)\overline{n} + \epsilon\left(\alpha y + \hat{\ell} - \hat{\sigma}\frac{\sigma}{1-\sigma}\right) - \beta\left(\alpha z - \hat{\ell}\frac{\ell}{1-\ell} - \hat{\sigma}\frac{\sigma}{1-\sigma}\right) \\
\implies &\hat{\delta} = (\epsilon-\beta)\left(\overline{n} - \hat{\sigma}\frac{\sigma}{1-\sigma}\right) + \alpha(\epsilon y - \beta z) + \hat{\ell}\left(\epsilon + \beta\frac{\ell}{1-\ell}\right)
\end{aligned} \tag{151}$$

Law of Motion: s

Next, I consider $\hat{s}$. Recall the FOC for s_t: $\frac{1-s_t}{s_t} = \frac{p_{bt}\dot{B}_t}{p_{at}\dot{A}_t} = \frac{p_{bt}((1-s_t)\sigma_t N_t)^\lambda B_t^\phi}{p_{at}(s_t\sigma_t N_t)^\lambda A_t^\phi}$. Taking logs and derivatives of both sides:

$$\begin{aligned}
&-\hat{s}\frac{s}{1-s} - \hat{s} = g_{p_bt} + \lambda\left(-\hat{s}\frac{s}{1-s} + \hat{\sigma} + \overline{n}\right) + \phi z - g_{p_at} - \lambda(\hat{s} + \hat{\sigma} + \overline{n}) - \phi y \\
\implies &\hat{s}\left(1 + \frac{s}{1-s}\right) = g_{p_at} - g_{p_bt} + \phi y - \phi z + \lambda\hat{s}\left(1 + \frac{s}{1-s}\right) \\
\implies &\hat{s}\frac{1-\lambda}{1-s} = g_{p_at} - g_{p_bt} + \phi y - \phi z
\end{aligned} \tag{152}$$

Recall FOCs for A_t and B_t: $\frac{\dot{p_{at}}}{p_{at}} = \rho - \frac{1}{p_{at}}[M_t u'(c_t)\alpha\frac{c_t}{A_t} + p_{at}\phi\frac{\dot{A}_t}{A_t} - \alpha\epsilon v_t M_t\frac{\delta_t}{A_t}]$ and $\frac{\dot{p_{bt}}}{p_{bt}} = \rho - \frac{1}{p_{bt}}[p_{bt}\phi\frac{\dot{B}_t}{B_t} + \alpha\beta v_t M_t\frac{\delta_t}{B_t}]$ respectively. Substituting, I get

$$\begin{aligned}\hat{s}\frac{1-\lambda}{1-s} &= \rho - \frac{1}{p_{at}}[M_t u'(c_t)\alpha\frac{c_t}{A_t} + p_{at}\phi\frac{\dot{A}_t}{A_t} - \alpha\epsilon v_t M_t\frac{\delta_t}{A_t}] \\ &\quad - \rho + \frac{1}{p_{bt}}[p_{bt}\phi\frac{\dot{B}_t}{B_t} + \alpha\beta v_t M_t\frac{\delta_t}{B_t}] + \phi y - \phi z \\ \implies \hat{s} &= \frac{1-s}{1-\lambda}\left[\frac{\alpha M_t\beta\delta_t v_t}{p_{bt}B_t} - \frac{\alpha M_t(u'(c_t)c_t - \epsilon\delta_t v_t)}{p_{at}A_t}\right] \end{aligned} \tag{153}$$

Recall the FOC for σ_t: $\frac{1-\sigma_t}{\sigma_t} = \frac{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}{\lambda(p_{at}\dot{A} + p_{bt}\dot{B})}$. From the FOC for s_t, I know $p_{bt}\dot{B}_t = \frac{1-s_t}{s_t}p_{at}\dot{B}_t$; substituting this yields:

$$\begin{aligned}\lambda\frac{1-\sigma_t}{\sigma_t} &= \frac{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}{p_{at}\dot{A}\left(1 + \frac{1-s_t}{s_t}\right)} \\ \implies \lambda\frac{y}{s_t}\frac{1-\sigma_t}{\sigma_t} &= \frac{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}{p_{at}A_t} \end{aligned} \tag{154}$$

Similarly, I get:

$$\lambda\frac{z}{1-s_t}\frac{1-\sigma_t}{\sigma_t} = \frac{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}{p_{bt}B_t} \tag{155}$$

Now, recall the FOC for ℓ_t: $\frac{1-\ell_t}{\ell_t} = \frac{\beta\delta_t\tilde{v}_t}{1-\epsilon\delta_t\tilde{v}_t}$. I manipulate this:

$$\begin{aligned} 1 - \ell_t - \epsilon\delta_t\tilde{v}_t + \ell_t\epsilon\delta_t\tilde{v}_t &= \ell_t\beta\delta_t\tilde{v}_t \\ \implies \ell_t(\beta-\epsilon)\delta_t v_t &= (1-\ell_t)u'(c_t)c_t - \epsilon\delta_t v_t \qquad (156) \\ \implies (1-\ell_t)u'(c_t)c_t &= \epsilon\delta_t v_t + \ell_t(\beta-\epsilon)\delta_t v_t \qquad (157) \end{aligned}$$

Combining (154) and (156) gives us

$$\lambda\frac{y}{s_t}\ell\frac{1-\sigma_t}{\sigma_t} = \frac{M_t[u'(c_t)c_t - \epsilon\delta_t v_t]}{p_{at}A_t} \tag{158}$$

Similarly, combining (155) and (157) gives us

$$\lambda\frac{z}{1-s_t}(1-\ell)\frac{1-\sigma_t}{\sigma_t} = \frac{M_t\beta\delta_t v_t}{p_{bt}B_t} \tag{159}$$

Substituting (158) and (159) into (153) yields:

$$\hat{s} = \frac{1-s}{1-\lambda}\left[\alpha\lambda\frac{z}{1-s_t}(1-\ell)\frac{1-\sigma_t}{\sigma_t} - \alpha\frac{y}{s_t}\ell\frac{1-\sigma_t}{\sigma_t}\right]$$
$$\implies \hat{s} = \alpha z\frac{\lambda}{1-\lambda}(1-\ell)\frac{1-\sigma}{\sigma} - \alpha y\frac{\lambda}{1-\lambda}\frac{1-s}{s}\ell\frac{1-\sigma}{\sigma} \tag{160}$$

Law of Motion: σ

I can use the FOC for σ_t, (157), (155) and rearrange:

$$\frac{1-\sigma_t}{\sigma_t} = \frac{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}{\lambda(p_{at}\dot{A} + p_{bt}\dot{B})}$$
$$\implies \frac{1-\sigma_t}{\sigma_t} = \frac{1-\ell}{1-\ell}\frac{M_t[u'(c_t)c_t + (\beta-\epsilon)\delta_t v_t]}{\lambda\left(1+\frac{s_t}{1-s_t}\right)(p_{bt}\dot{B})}$$
$$\implies \frac{1-\sigma_t}{\sigma_t} = \frac{1-s_t}{1-\ell}\frac{M_t[\beta\delta_t v_t]}{\lambda(p_{bt}g_{Bt}B_t)} \tag{161}$$

Taking logs and derivatives yields:

$$-\hat{\sigma}\frac{\sigma}{1-\sigma} - \hat{\sigma} = -\hat{s}\frac{s}{1-s} + \hat{\ell}\frac{\ell}{1-\ell} - \delta + \hat{\delta} + g_{vt} - g_{p_b t} - \hat{z} - z \tag{162}$$

From the FOC for M_t, I get $g_{vt} = \frac{\dot{v}_t}{v_t} = \rho - \frac{u(c_t)}{v_t} + \delta_t$. From the FOC for B_t I get $g_{p_b t} = \frac{\dot{p_{bt}}}{p_{bt}} = \rho - \frac{1}{p_{bt}}[p_{bt}\phi\frac{\dot{B}_t}{B_t} + \alpha\beta v_t M_t\frac{\delta_t}{B_t}]$. I can substitute in these expressions and the expressions I previously found for $\hat{\delta}$ and $\hat{z}$. This yields:

$$-\hat{\sigma}\frac{\sigma}{1-\sigma} - \hat{\sigma} = -\hat{s}\frac{s}{1-s} + \hat{\ell}\frac{\ell}{1-\ell} - \delta + \hat{\delta} + \rho - \frac{u(c_t)}{v_t} + \delta_t - \rho + \phi z + \frac{\alpha M_t\beta\delta_t v_t}{p_{bt}B_t} - \hat{z} - z$$
$$\implies -\hat{\sigma}\frac{1}{1-\sigma} = -\hat{s}\frac{s}{1-s} + \hat{\ell}\frac{\ell}{1-\ell} + (\epsilon-\beta)\left(\overline{n} - \hat{\sigma}\frac{\sigma}{1-\sigma}\right) + \alpha(\epsilon y - \beta z) + \hat{\ell}\left(\epsilon + \beta\frac{\ell}{1-\ell}\right)$$
$$- \frac{u(c_t)}{v_t} + (\phi-1)z - \lambda\left(\overline{n} + \hat{\sigma} - \hat{s}\frac{s}{1-s}\right) + (1-\phi)z + \frac{\alpha M_t\beta\delta_t v_t}{p_{bt}B_t}$$
$$\implies -\hat{\sigma}\left(\frac{1}{1-\sigma} + (\beta-\epsilon)\frac{\sigma}{1-\sigma} - \lambda\right) = -\hat{s}\frac{s}{1-s}(1-\lambda) + \hat{\ell}\left(\frac{\ell}{1-\ell} + \epsilon + \beta\frac{\ell}{1-\ell}\right)$$
$$+ \overline{n}(\epsilon - \beta - \lambda) + \alpha\epsilon y - \alpha\beta z - \frac{u(c_t)}{v_t} + \frac{\alpha M_t\beta\delta_t v_t}{p_{bt}B_t} \tag{163}$$

We can plug in (159) and rearrange:

$$\implies \hat{\sigma}\frac{1+(\beta-\epsilon)\sigma-\lambda(1-\sigma)}{1-\sigma} = (1-\lambda)\frac{s}{1-s}\hat{s} + (\lambda+\beta-\epsilon)\overline{n} + \alpha\beta z - \alpha\epsilon y + \frac{u(c_t)}{v_t} + (-\left(\frac{\ell}{1-\ell}(1+\beta)+\epsilon\right))\hat{\ell} - \alpha\lambda z\frac{1-\ell}{1-s_t}\frac{1-\sigma_t}{\sigma_t}$$

I set the following definitions:

$$\theta_\sigma = \frac{1-\sigma}{1+(\beta-\epsilon)\sigma-\lambda(1-\sigma)} \tag{164}$$

$$\omega_\sigma = -\left(\frac{\ell}{1-\ell}(1+\beta)+\epsilon\right) \tag{165}$$

$$\text{Ⓑ} = (1-\lambda)\frac{s}{1-s}\hat{s} + (\lambda+\beta-\epsilon)\overline{n} + \alpha\beta z - \alpha\epsilon y + \frac{u(c_t)}{v_t} - \alpha\lambda z\frac{1-\ell}{1-s_t}\frac{1-\sigma_t}{\sigma_t} \tag{166}$$

Then:

$$\hat{\sigma} = \theta_\sigma(\text{Ⓑ} + \omega_\sigma\hat{\ell}) \tag{167}$$

Law of Motion: ℓ

Recall the FOC for ℓ_t: $\frac{1-\ell_t}{\ell_t} = \frac{\beta\delta_t\frac{v_t}{u'(c_t)c_t}}{1-\epsilon\delta_t\frac{v_t}{u'(c_t)c_t}}$. Since $u'(c_t)c_t = c^{1-\gamma}$, taking logs and derivatives, this yields:

$$-\hat{\ell}\frac{1}{1-\ell} = \hat{\delta} + g_{vt} - (1-\gamma)g_{ct} + \frac{\epsilon\delta_t\frac{v_t}{u'(c_t)c_t}}{1-\epsilon\delta_t\frac{v_t}{u'(c_t)c_t}}\left(\hat{\delta} + g_{vt} - (1-\gamma)g_{ct}\right)$$

Substituting $\frac{1-\ell_t}{\ell_t}\frac{\epsilon}{\beta} = \frac{\epsilon\delta_t\frac{v_t}{u'(c_t)c_t}}{1-\epsilon\delta_t\frac{v_t}{u'(c_t)c_t}}$ I get:

$$\hat{\ell}\frac{1}{1-\ell} = -\left(1+\frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)(\hat{\delta} + g_{vt} + (\gamma-1)g_{ct}) \tag{168}$$

From the FOC for M_t, I get $g_{vt} = \frac{\dot{v}_t}{v_t} = \rho - \frac{u(c_t)}{v_t} + \delta_t$. It follows directly

that $g_{ct} = \alpha y + \hat{\ell} - \hat{\sigma}\frac{\sigma}{1-\sigma}$. I can substitute in these two expressions and $\hat{\delta}$:

$$\begin{aligned}
\hat{\ell} = &-(1-\ell)\left(1+\frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)\cdot \\
&[(\epsilon-\beta)\left(\overline{n}-\hat{\sigma}\frac{\sigma}{1-\sigma}\right)+\alpha(\epsilon y-\beta z) \\
&+\hat{\ell}\left(\epsilon+\beta\frac{\ell}{1-\ell}\right)+\rho-\frac{u(c_t)}{v_t}+\delta_t+(\gamma-1)\left(\alpha y+\hat{\ell}-\hat{\sigma}\frac{\sigma}{1-\sigma}\right)] \\
\\
\implies &\hat{\ell}\left[1+(\gamma-1+\epsilon+\beta\frac{\ell}{1-\ell})(1-\ell)\left(1+\frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)\right] \\
&=(1-\ell)\left(1+\frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)\cdot[(\gamma-1+\epsilon-\beta)\frac{\sigma}{1-\sigma}\hat{\sigma}+(\beta-\epsilon)\overline{n} \\
&+(1-\gamma-\epsilon)\alpha y+\alpha\beta z-\rho-\delta+\frac{u(c_t)}{v_t}] \qquad (169)
\end{aligned}$$

I set the following definitions:

$$\theta_\ell = \frac{(1-\ell)\left(1+\frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)}{1+\left(\gamma-1+\epsilon+\beta\frac{\ell}{1-\ell}\right)(1-\ell)\left(1+\frac{1-\ell}{\ell}\frac{\epsilon}{\beta}\right)} \qquad (170)$$

$$\omega_\ell = (\gamma-1+\epsilon-\beta)\frac{\sigma}{1-\sigma} \qquad (171)$$

$$Ⓐ = (\beta-\epsilon)\overline{n}+(1-\gamma-\epsilon)\alpha y+\alpha\beta z-\rho-\delta+\frac{u(c_t)}{v_t} \qquad (172)$$

Then:

$$\hat{\ell} = \theta_\ell(Ⓐ+\omega_\ell\hat{\sigma}) \qquad (173)$$

I substitute in (167):

$$\begin{aligned}
\hat{\ell} &= \theta_\ell(Ⓐ+\omega_\ell\theta_\sigma(Ⓑ+\omega_\sigma\hat{\ell})) \\
\implies \hat{\ell}[1-\omega_\ell\omega_\sigma\theta_\sigma] &= \theta_\ell(Ⓐ+\omega_\ell\theta_\sigma Ⓑ) \\
\implies \hat{\ell} &= \frac{\theta_\ell(Ⓐ+\omega_\ell\theta_\sigma Ⓑ)}{1-\omega_\ell\omega_\sigma\theta_\sigma} \qquad (174)
\end{aligned}$$

Addendum: $\frac{u(c_t)}{v_t}$

To determine both Ⓐ and Ⓑ, I need an expression for $\frac{u(c_t)}{v_t}$. First, recall the FOC for ℓ_t: $\frac{1-\ell_t}{\ell_t} = \frac{\beta\delta_t \frac{v_t}{u'(c_t)c_t}}{1-\epsilon\delta_t \frac{v_t}{u'(c_t)c_t}}$. Thus,

$$\begin{aligned}
\frac{\ell}{1-\ell}\beta\delta\frac{u(c_t)}{u'(c_t)c_t} &= \frac{u(c_t)}{v_t}(1-\epsilon\delta\frac{v_t}{u'(c_t)c_t}) \\
\implies \frac{\ell}{1-\ell}\beta\delta\frac{u(c_t)}{u'(c_t)c_t} &= \frac{u(c_t)}{v_t} - \epsilon\delta\frac{u(c_t)}{u'(c_t)c_t} \\
\implies \frac{\ell}{1-\ell}\beta\delta\frac{u(c_t)}{u'(c_t)c_t} + \epsilon\delta\frac{u(c_t)}{u'(c_t)c_t} &= \frac{u(c_t)}{v_t} \qquad (175)
\end{aligned}$$

Let $\tilde{u} = \frac{u(c)}{u'(c)c}$. Then,

$$\tilde{u} = \frac{u(c_t)}{u'(c_t)c_t} = \overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma} \qquad (176)$$

Thus, I need an expression for c_t. Look at δ:

$$\begin{aligned}
\delta_t &= \overline{\delta}N_t^{\epsilon-\beta}c_t^{\epsilon}h_t^{-\beta} \\
\implies \frac{\delta_t}{\overline{\delta}}\left(\frac{c_t}{h_t}\right)^{-\beta} &= N_t^{\epsilon-\beta}c_t^{\epsilon}h_t^{-\beta}\frac{c_t^{-\beta}}{h_t^{-\beta}} \\
\implies \left(\frac{\delta_t}{\overline{\delta}}\left(\frac{c_t}{h_t}\right)^{-\beta}\right)^{\frac{1}{\epsilon-\beta}} &= N_t c_t \\
\implies \frac{\left(\frac{\delta_t}{\overline{\delta}}\left(\frac{c_t}{h_t}\right)^{-\beta}\right)^{\frac{1}{\epsilon-\beta}}}{N_t} &= c_t \qquad (177)
\end{aligned}$$

Next, note that

$$\begin{aligned}
\frac{c_t}{h_t} &= \frac{A^{\alpha}\ell_t(1-\sigma_t)}{B^{\alpha}(1-\ell_t)(1-\sigma_t)} \\
\implies \frac{c_t}{h_t} &= \frac{\ell_t}{1-\ell_t}\left(\frac{A}{B}\right)^{\alpha}
\end{aligned}$$

Finally, note that

$$\begin{aligned}
\frac{z}{y} &= \frac{\dot{B}_t A_t}{B_t \dot{A}_t} = \frac{A_t (1-s_t)^{\lambda} \sigma_t^{\lambda} N_t^{\lambda} B_t^{\phi}}{B_t s_t^{\lambda} \sigma_t^{\lambda} N_t^{\lambda} A_t^{\phi}} \\
\implies \frac{z}{y} &= \frac{A_t^{1-\phi}}{B_t^{1-\phi}} \left(\frac{1-s_t}{s_t} \right)^{\lambda} \\
\implies \left(\frac{z}{y} \right)^{\frac{1}{1-\phi}} &= \frac{A_t}{B_t} \left(\frac{1-s_t}{s_t} \right)^{\frac{\lambda}{1-\phi}} \\
\implies \left(\frac{z}{y} \right)^{\frac{1}{1-\phi}} &= \frac{A_t}{B_t} \left(\frac{1-s_t}{s_t} \right)^{\frac{\lambda}{1-\phi}} \\
\implies \left(\frac{z}{y} \right)^{\frac{1}{1-\phi}} \left(\frac{s_t}{1-s_t} \right)^{\frac{\lambda}{1-\phi}} &= \frac{A_t}{B_t} \qquad (178)
\end{aligned}$$

Thus:

$$c_t = \frac{\left(\frac{\delta_t}{\bar{\delta}} \left(\frac{\ell_t}{1-\ell_t} \left(\left(\frac{z}{y} \right)^{\frac{1}{1-\phi}} \left(\frac{s_t}{1-s_t} \right)^{\frac{\lambda}{1-\phi}} \right)^{\alpha} \right)^{-\beta} \right)^{\frac{1}{\epsilon-\beta}}}{N_t} \qquad (179)$$

A.11 Proof of Proposition 10

We know from the FOCs of the Hamiltonian that:

$$\tilde{v}_t = \frac{\tilde{u}_t}{\rho - \delta_t + g_{vt}}, \quad \tilde{u}_t = \frac{u(c_t)}{u'(c_t)c_t} \tag{180}$$

First, consider the denominator of $\tilde{v}_t$: $\rho - \delta_t + g_{vt}$. As $\delta_t \to 0$, $\rho - \delta_t + g_{vt} \to \rho + g_{vt}$.

Next, consider g_{vt}. We know from our FOCs that

$$v_t = \frac{u(c_t)}{\rho - \delta_t + g_{vt}}$$

If the denominator converges to a constant as $\delta_t \to 0$ and $\tilde{u} \to \infty$, we get

$$g_{vt} \to \frac{\dot{u}(c_t)}{u(c_t)} = \frac{u'(c_t)\dot{c}}{u(c_t)} = \frac{u'(c_t)cg_{ct}}{u(c_t)} \tag{181}$$

$$\implies g_{vt} \to \frac{1}{\tilde{u}_t} g_{ct} \tag{182}$$

This means $g_{vt} \to 0$ as $\tilde{u} \to \infty$, so indeed means the denominator v_t converges to a constant.

This implies that as $\delta_t \to 0$ and $\tilde{u} \to \infty$:

$$\tilde{v}_t = \frac{\tilde{u}_t}{\rho - \delta_t + g_{vt}} \to \frac{\tilde{u}_t}{\rho} \tag{183}$$

This immediately implies

$$E_\rho^{\tilde{v}} \to -1 \tag{184}$$

and

$$E_c^{\tilde{v}} \to E_c^{\tilde{u}} \tag{185}$$

We now turn to $E_c^{\tilde{u}}$. We know

$$\tilde{u}(c_t) = \overline{u} c_t^{\gamma-1} + \frac{1}{1-\gamma}$$

$$\implies \tilde{u}'(c_t) = (\gamma - 1)\overline{u} c^{\gamma-2} \tag{186}$$

Thus,

$$E_c^{\tilde{u}} = \frac{\tilde{u}'(c_t)c_t}{\tilde{u}(c_t)} = \frac{(\gamma-1)\overline{u}c_t^{\gamma-1}}{\overline{u}c_t^{\gamma-1} + \frac{1}{1-\gamma}} \tag{187}$$

$$\implies E_c^{\tilde{u}} = \frac{(\gamma-1)\left(\tilde{u}_t + \frac{1}{\gamma-1}\right)}{\tilde{u}} \tag{188}$$

$$\implies E_c^{\tilde{u}} \to (\gamma-1) \tag{189}$$

$$\implies E_c^{\tilde{v}} \to (\gamma-1) \tag{190}$$

as $\tilde{u}_t \to \infty$.

Finally, to calculate the ρ' that increases $\tilde{v}_t$ equivalent to a doubling of consumption (starting from $\rho^*\%$) for the table in the main text, I find the ρ' that satisfies:

$$2^{E_c^{\tilde{v}}} = \left(\frac{\rho'}{\rho^*}\right)^{E_c^{\tilde{v}}} \tag{191}$$

$$\implies 2^{(\gamma-1)} = \left(\frac{\rho'}{\rho^*}\right)^{-1} \tag{192}$$

$$\implies \rho' = 2^{(1-\gamma)}\rho^* \tag{193}$$

B Numerical Simulation

B.1 Simulating the Transition Dynamics

I solve the system of differential equations characterizing the optimal allocation numerically using "reverse shooting" (like Jones (2016)). I start from the steady state, consider a small deviation, and then run time backwards. In the notation that follows, I start from time T and run time backwards to time 0.

Given values for the parameters γ, ϵ, β, ρ, λ, ϕ, α, $\overline{n}$, $\overline{u}$, and $\overline{\delta}$, as well as a specified N_T and a small $\delta_T > 0$ (small deviation from the steady state), I need to find values of s_T and ℓ_T. To do this, I use the function 'fminsearch' in Matlab to find values of s_T and ℓ_T that minimize the distance between $\hat{s}_T$, $\hat{\ell}_T$ and $\hat{\sigma}_T$ and their steady state values. I then run time backwards, giving us a candidate path.

To determine the values for the other parameters, I first pick ϵ, β, and γ. I also set ϕ Then, I use the function 'patternsearch' in Matlab to find values for λ, $\overline{\delta}$, $\overline{u}$, N_T and δ_T which minimize the weighted sum of the deviations from a selection of moments of the candidate path and a set of preferred values. These moments are given below.

1. Given a candidate path, I first find the year t_0 in which $\tilde{u}$, the value of a year of life as a ratio to consumption, is closest to 4. The first moment is $\tilde{u}_{t_0}$ compared to 4.

2. The second moment is the growth rate of consumption at t_0 compared to 1 percent.

3. The third moment is ℓ_{t_0}, the fraction of workers in the consumption sector, compared to 95%.

4. The fourth moment is the growth rate of proportion of the population working as scientists at time t_0, $g_{\sigma t_0}$, compared to 2%.

5. The fifth moment is hazard rate δ at time t_0, compared to 0.1%.

6. The sixth moment is the growth rate of δ at time T compared to g_δ, to ensure the the simulation is close to steady state at time T.

I pick $\gamma = 1.5$, $\epsilon = 0.4$, and $\beta = 0.3$ as reasonable parameters. In addition, I set $\rho = 0.02$, $\alpha = 1$, and $\overline{n} = 1\%$. Note that these choices don't seem to matter for the qualitative results (as long as $\epsilon > \beta$ and $\gamma > 1$). The process described above can find different local minima depending on the initial guess as well the value of ϕ supplied, so I hunt for the best overall fit. I end up using $\phi = 5/6$ and $\lambda = 0.3$, $\overline{\delta} = 3.8965 \times 10^{-5}$, $\overline{u} = 0.0098$,

$N_T = 9.2955 \times 10^{14}$,and $\delta_T = 5 \times 10^{-4}$.

To extrapolate M_∞, I have to calculate the area under the hazard rate curve, i.e. $\int_{t_0}^{\infty} \delta_s ds$. Note that:

$$\begin{aligned} & \int_{t_0}^{\infty} \delta_s ds = \int_{t_0}^{T} \delta_s ds + \int_{T}^{\infty} \delta_s ds \\ \implies & \int_{t_0}^{\infty} \delta_s ds \approx \int_{t_0}^{T} \delta_s ds + \int_{0}^{\infty} \delta_T \cdot e^{-s g_\delta} ds \\ \implies & \int_{t_0}^{\infty} \delta_s ds \approx \int_{t_0}^{T} \delta_s ds + \frac{\delta_T}{g_\delta} \end{aligned} \tag{194}$$

since at time T, we are approximately at the steady state, where δ declines exponentially.

Thus, I sum the area under our simulated δ from the time representing today to the end of the simulation, which gives us $\int_{t_0}^{T} \delta_s ds$. Verifying that indeed $\hat{\delta}_T \approx g_\delta$, I can then calculate $\frac{\delta_T}{g_\delta}$. Summing these two terms gives us the desired $\int_{t_0}^{\infty} \delta_s ds$, and then the probability of humanity surviving to infinity conditional on surviving until t_0 is then $e^{-\int_{t_0}^{\infty} \delta_s ds}$.

B.2 Simulating the Acceleration in Growth

The natural way to simulate the acceleration of growth (in this case, faster population growth) would be to solve the differential equations characterizing the optimal allocation using "forward shooting". However, due to the instability of the system of differential equations, this yields unreliable results. Thus, I again proceed by solving the system of differential equations using "reverse shooting" as when we simulated the transition dynamics.

First, the transition path without the acceleration in growth is given by the path as found in Appendix B.1. I will refer to this as the unperturbed path.

Next, we would like to simulate the transition path with accelerated growth. I use the same parameters as in Appendix B.1 except for a time-varying rate of $\overline{n}$, set as discussed in the main text. This gives me a candidate path with accelerated growth. I would like to find a transition path with acceleration that matches the unperturbed path up until the moment of acceleration. Thus, using the function 'fminsearch' in Matlab, I find δ_T and N_T that yield a candidate path with accelerated growth that minimizes

the weighted sum of the deviations from a selection of moments and a set of preferred values.

In particular, I pick some year t_0 prior to the year in which growth accelerates; this will be the reference year on the unperturbed path. Given a candidate path, I find a year t^* in which δ_{t^*} of our candidate path is closest to δ_{t_0} on the unperturbed path. Then, my moments are s_{t^*}, ℓ_{t^*}, σ_{t^*}, δ_{t^*}, y_{t^*}, z_{t^*}, and N_{t^*}, compared to their respective values at t_0 on the unperturbed transition path. Since s, ℓ, σ, δ, y, z, and N uniquely characterize all the variables of our economy on the optimal allocation and both the unperturbed and the accelerated path evolve according to the same system of differential equations prior to the acceleration, this ensures that both the unperturbed and accelerated transition path represent the same economy up until the moment where growth is accelerated.

I experiment with the weights and the reference year to hunt for the best overall fit. I end up picking $\delta_T = 5.0326 \times 10^{-4}$ and $N_T = 9.3991 \times 10^{14}$ for the accelerated transition path that results in a permanent level effect, and $\delta_T = 5.0001 \times 10^{-4}$ and $N_T = 9.2948 \times 10^{14}$ for the transition path with the temporary boom.

This method appears to work well (i.e. matches the unperturbed path very well) for an acceleration in growth that is not too large, although it is still imperfect. However, it enables us to sidestep the difficulty of "forward shooting" and compare the transition paths with and without an acceleration in growth using "reverse shooting".

I extrapolate the long-term survival probability M as before.

References

Acemoglu, Daron, "Directed Technical Change," *Review of Economic Studies*, 2002, *69* (4), 781–809.

__ , Philippe Aghion, Leonardo Bursztyn, and David Hemous, "The Environment and Directed Technical Change," *American Economic Review*, February 2012, *102* (1), 131–166.

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

Bloom, Nicholas, Charles Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?," Technical Report w23782, National Bureau of Economic Research, Cambridge, MA September 2017.

Bostrom, Nick, "Existential Risks: Analyzing Human Extinction Scenarios," *Journal of Evolution and Technology, Vol. 9, No. 1 (2002)*, March 2002, *9*, 1–35.

__ , "Astronomical Waste: The Opportunity Cost of Delayed Technological Development," *Utilitas*, November 2003, *15* (3), 308–314.

Brock, William and M. Scott Taylor, "Economic Growth and the Environment: A Review of Theory and Empirics," Handbook of Economic Growth, Elsevier 2005.

Caplan, Bryan, "The Totalitarian Threat," in "Global Catastrophic Risks" 2008, p. 498.

Chetty, Raj, "A New Method of Estimating Risk Aversion," *American Economic Review*, December 2006, *96* (5), 1821–1834.

Farquhar, Sebastian, John Halstead, Owen Cotton-Barratt, Stefan Schubert, Hadyn Belfield, and Andrew Snyder-Beattie, "Existential Risk: Diplomacy and Governance," Technical Report, Global Priorities Project, Oxford University 2017.

Hall, Robert E., "Reconciling Cyclical Movements in the Marginal Value of Time and the Marginal Product of Labor," *Journal of Political Economy*, April 2009, *117* (2), 281–323.

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

Jones, Charles I., "R & D-Based Models of Economic Growth," *Journal of Political Economy*, 1995, *103* (4), 759–784.

_ , "Life and Growth," *Journal of Political Economy*, March 2016, *124* (2), 539–578.

_ and Paul M. Romer, "The New Kaldor Facts: Ideas, Institutions, Population, and Human Capital," *American Economic Journal: Macroeconomics*, January 2010, *2* (1), 224–245.

Lucas, Deborah J., "Asset Pricing with Undiversifiable Income Risk and Short Sales Constraints: Deepening the Equity Premium Puzzle," *Journal of Monetary Economics*, December 1994, *34* (3), 325–341.

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

_ and Robert S Pindyck, "Welfare Costs of Catastrophes: Lost Consumption and Lost Lives," Working Paper 26068, National Bureau of Economic Research July 2019.

Méjean, Aurélie, Antonin Pottier, Stéphane Zuber, and Marc Fluerbaey, "Intergenerational Equity under Catastrophic Climate Change," Technical Report 2017.25, FAERE - French Association of Environmental and Resource Economists November 2017.

_ , _ , _ , and _ , "When Opposites Attract: Averting a Climate Catastrophe despite Differing Ethical Views," Technical Report 2019.

Nordhaus, William and Paul Sztorc, "DICE 2013R: Introduction and User's Manual," October 2013.

Parfit, Derek, *On What Matters: Volume Two*, Oxford University Press, May 2011.

Pindyck, Robert S., "Climate Change Policy: What Do the Models Tell Us?," *Journal of Economic Literature*, September 2013, *51* (3), 860–872.

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

Romer, Paul, "Cake Eating, Chattering, and Jumps: Existence Results for Variational Problems," *Econometrica*, 1986, *54* (4), 897–908.

__, "Endogenous Technological Change," *Journal of Political Economy*, 1990, *98* (5), S71–102.

Sagan, Carl, *Pale Blue Dot: A Vision of the Human Future in Space*, Random House Publishing Group, 1994.

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

Solow, Robert M., "A Contribution to the Theory of Economic Growth," *The Quarterly Journal of Economics*, 1956, *70* (1), 65–94.

Stern, Nicholas, "The Stern Review on the Economic Effects of Climate Change," *Population and Development Review*, 2006, *32* (4), 793–798.

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

Torres, Phil, "How Likely Is an Existential Catastrophe?," September 2016.

Weitzman, Martin L., "Subjective Expectations and Asset-Return Puzzles," *American Economic Review*, September 2007, *97* (4), 1102–1130.