The A.I. Dilemma: Growth versus Existential Risk

Charles I. Jones*
Stanford GSB and NBER

June 13, 2023 — Version 0.5
Preliminary, comments appreciated

Abstract

Advances in artificial intelligence (A.I.) are a double-edged sword. On the one hand, they may increase economic growth as A.I. augments our ability to innovate or even itself learns to discover new ideas. On the other hand, many experts note that these advances entail existential risk: creating a superintelligent entity misaligned with human values could lead to catastrophic outcomes, including human extinction. This paper considers the optimal use of A.I. technology in the presence of these opportunities and risks. Under what conditions should we continue the rapid progress of A.I. and under what conditions should we stop?

*I'm grateful to Jean-Felix Brouillette, Tom Davidson, Sebastian Di Tella, Pete Klenow, Anton Korinek, Pascual Restrepo, Charlotte Siegmann, Chris Tonetti, and Phil Trammell for helpful comments and discussions.

1. Introduction

Recent advances in artificial intelligence (A.I.) will surely raise living standards in the coming years. Protein-folding, speech recognition, and the amazing accomplishments of generative models in producing text and images have sped past expectations from just a few years ago (Bubeck et al., 2023). It seems likely that A.I. will augment our ablities to innovate in the near term, and it is certainly within the realm of possibility that A.I. could exceed human intelligence at many cognitive tasks and even begin innovating itself. Once machines can produce ideas, the limits to growth set by the quantity and quality of researchers may no longer hold, and growth rates could speed up, potentially even leading to a singularity with infinite consumption. Models along these lines have been explored by Aghion, Jones and Jones (2019), Trammell and Korinek (2020), and Davidson (2021).

On the other hand, as many experts have warned recently, these advances do not come without risk. A substantial contingent of the A.I. research community, including leading researchers at OpenAI and Google, warn that these advances could constitute an existential risk for humanity: if we create an intelligence smarter than humans that is not aligned with our goals, there is some risk that humans could be left behind or even annihilated. These concerns raise substantial questions about whether we should pause our research on A.I. or perhaps stop it altogether at some point.

More succinctly, A.I. could raise living standards by more than electricity or the internet. But it may pose risks that are exceed those from nuclear weapons. This paper considers the optimal use of A.I. in the presence of this double-edged sword. Under what conditions should we continue the rapid progress of A.I. and under what conditions should we stop?

The goal of the paper is not to provide an exact answer to this question, as the answer will surely depend on parameters that we cannot precisely quantify. Instead, the paper develops some simple models to elucidate the economic forces that are involved in thinking through these questions.

Several insights emerge:

1. The curvature of utility is very important. With log utility, the models are remarkably unconcerned with existential risk, suggesting that large consumption gains that A.I. might deliver can be worth gambles that involve a 1-in-3 chance of

extinction.

2. For CRRA utility with a risk aversion coefficient (γ) of 2 or more, the picture changes sharply. These utility functions are bounded, and the marginal utility of consumption falls rapidly. Models with this feature are quite conservative in trading off consumption gains versus existential risk.

3. These findings even extend to singularity scenarios. If utility is bounded — as it is in the standard utility functions we use frequently in a variety of applications in economics — then even infinite consumption generates relatively small gains. The models with bounded utility remain conservative even when a singularity delivers infinite consumption.

4. A key exception to this conservative view of existential risk emerges if the rapid innovation associated with A.I. leads to new technologies that extend life expectancy and reduce mortality. These gains are "in the same units" as existential risk and do not run into the sharply declining marginal utility of consumption. Even with a future-oriented focus that comes from low discounting, A.I.-induced mortality reductions can make large existential risks bearable.

In Section 2, we develop a simple model to illustrate some of these forces as clearly as possible. Section 3 then extends the analysis to include a richer theory of dynamics, the possibility of a singularity, and the prospect that the innovations from A.I. could also extend life expectancy.

Related literature. Serious concerns about the existential risk associated with artificial intelligence have been highlighted in recent decades by Joy (2000), Bostrom (2002, 2014), Rees (2003), Posner (2004), and Yudkowsky et al. (2008). These concerns have accelerated together with the progress of A.I. itself. Ngo, Chan and Mindermann (2023) provides a recent overview of how these concerns could materialize in the context of the "alignment problem."

Jones (2016) considers the tradeoffs between the economic benefits of new technologies and their potential costs in terms of lost lives, for example because of nuclear weapons, biohazards, or even the risks associated with frontier science. That paper argues that as we get richer, it may be optimal to slow the rate of economic growth, or at

least redirect innovation toward life-saving technologies. This paper differs by focusing explicitly on the amazing potential benefits as well as the existential risk associated with artificial intelligence.

Aschenbrenner (2020) extends the framework in Jones (2016) to focus on existential risk, positing that existential risk is increasing in aggregate consumption and decreasing in aggregate mitigation efforts. He suggests we may live in a critical "time of perils" in which we are advanced enough to face high risk but not rich enough to spend sufficiently on mitigation efforts. Martin and Pindyck (2015, 2020) consider catastrophes and how the value of statistical life (VSL) can be used to evaluate the gains from avoiding catastrophes. All of this work — as well as the present paper — builds on Rosen (1988), Murphy and Topel (2003), Nordhaus (2003), and Hall and Jones (2007) in thinking about how to value lives.

2. A Simple Model

For the first model, suppose that advances in A.I. allow computers to augment and even substitute for humans in innovation, leading to an acceleration of economic growth to some rate g, say something like 10% per year. However, the use of this A.I. poses an existential risk to humanity. Using the advanced A.I. for T periods leads to a consumption per person of $c_T = c_0 e^{gT}$, but at the same time, the probability that the world survives is $S(T) = e^{-\delta T}$.

We simplify further so that the model is essentially static. The only decision is to choose T, the intensity of using the A.I. All growth and existential risk is realized immediately rather than over time, and if society survives, people consume the constant c_T forever after.

Social welfare for a constant population of N people getting the constant flow utility $u(c)$ forever is

$$U = N \int_0^\infty e^{-\rho t} u(c) dt = \frac{1}{\rho} N u(c).$$

Constant exogenous rates of population growth or decline would only change the discount rate ρ to $\rho - n$; it is therefore already included implicitly.

The setup then reduces to the static problem of choosing T to maximize expected

utility, where the expectation is taken with respect to existential risk:

$$EU = S(T) \cdot \frac{1}{\rho} N u(c) = e^{-\delta T} \cdot \frac{1}{\rho} N u(c_0 e^{gT}).$$

Notice that the N and the ρ just scale up or down social welfare but will drop out of the first order condition. The N people each benefit from the higher growth and each suffer the loss if the world ends. And the present value of the infinite future is simply proportional to the annual flow $u(c)$ via $1/\rho$. Also, we've normalized the utility of death to zero, an assumption we will discuss later.

The first-order condition for optimality is

$$\begin{aligned} S'(T)u(c) + S(T)u'(c)\frac{dc}{dT} &= 0 \\ \Rightarrow S'(T)u(c) + S(T)u'(c)c\frac{d\log c}{dT} &= 0 \\ \Rightarrow v(c) \equiv \frac{u(c)}{u'(c)c} = \frac{d\log c/dT}{-d\log S/dT} &= \frac{g}{\delta} \end{aligned}$$

Writing this more succinctly, the solution is to choose T^* such that $c^* = c_0 e^{gT^*}$ satisfies

$$\boxed{v(c^*) = \frac{g}{\delta}} \tag{1}$$

The left side of this equation is $v(c) \equiv u(c)/u'(c)c$, which is the value of a year of life measured in years of consumption. For example, in the United States today, a typical value of a year life is around $250,000, which comes from VSL's of around $10 million for a 40-year old who might live for 40 more years. Because consumption per person is around $40,000, this value of life implies $v(c_{us,today}) \approx 6$. That is, a year of life is worth six times per capita consumption. I like to think of the units of $v(c)$ as this: how much would you pay, measured in years worth of per capita consumption, for one more year of life?

One way of understanding this first-order condition is to note that it is optimal to use the A.I. as long as

$$\underset{\text{Lost lives}}{\delta\, v(c)} \leq \underset{\text{Extra growth}}{g}$$

If you let the A.I. run for one more period, the cost is a probability δ of ending the world, which is a loss of $v(c)$ per person. The benefit is the extra period of consumption growth

at rate g. The optimal choice of how long to let the A.I. run equates the cost and benefit at the margin.

2.1 CRRA Utility

Equation (1) implicitly defines the optimal level of consumption c^*. We choose the amount of time T^* to let the A.I. run until the value of life is equal to g/δ, which I think of as the "A.I. Benefit-Cost" ratio, or AIBC ratio for short.

To solve further, we assume the CRRA functional form for utility:

$$u(c) = \begin{cases} \bar{u} + \frac{c^{1-\gamma}}{1-\gamma} & \text{if } \gamma \neq 1 \\ \bar{u} + \log c & \text{if } \gamma = 1 \end{cases}$$

With CRRA utility, the value of life is given by

$$v(c) \equiv \frac{u(c)}{u'(c)c} = \begin{cases} \bar{u}c^{\gamma-1} + \frac{1}{1-\gamma} & \text{if } \gamma \neq 1 \\ \bar{u} + \log c & \text{if } \gamma = 1 \end{cases} \tag{2}$$

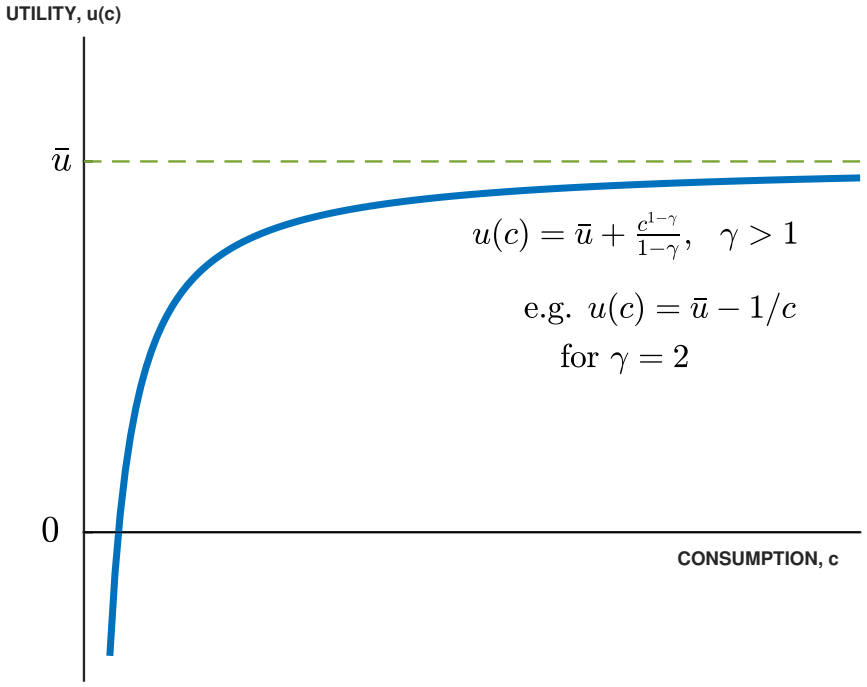
We will focus on the cases of $\gamma > 1$ and $\gamma = 1$ (log utility) as being most relevant. A large literature in macroeconomics focuses on these cases. However, it will be easy to see what happens if $\gamma < 1$.

Crucially, notice that the value of life $v(c)$ rises with consumption for $\gamma \geq 1$. To see the intuition for this fact, consider Figure 1 and note that utility is bounded for $\gamma > 1$. In this case, the marginal utility of consumption falls rapidly, and flow utility can never be larger than the parameter $\bar{u}$.

You can also see from Figure 1 why the parameter $\bar{u}$ is important. In setting up the problem, we normalized the utility when dead to zero; this is a free normalization and we could have chosen any other value. However, once death is zero, life must give positive utility in order to be preferred. With $\gamma > 1$, the term $c^{1-\gamma}/(1-\gamma)$ is less than zero. In other words, unless we do something — like adding a constant $\bar{u} > 0$ — life would not be preferred to death.

With $\gamma > 1$, the parameter $\bar{u}$ can be interpreted as the maximum flow utility an individual can obtain, even with infinite consumption. Put differently, even with infinite consumption, the utility from living remains finite. This has important implications

Figure 1: Bounded flow utility when $\gamma > 1$



Note: For $\gamma > 1$, CRRA utility is bounded, and the upper bound is given by the parameter $\bar{u}$.

for A.I. Even a singularity that delivers infinite consumption sometime in the next few decades only leads to a flow utility of $\bar{u}$, not to infinite utility. In contrast, if $\gamma < 1$ or even in the log case of $\gamma = 1$, then infinite consumption would lead to infinite utility.

Combining the equation for $v(c)$ in (2) with the key first-order condition in (1) gives

$$c^* = \begin{cases} \left[\frac{1}{\bar{u}}\left(\frac{g}{\delta} + \frac{1}{\gamma-1}\right)\right]^{\frac{1}{\gamma-1}} & \text{if } \gamma > 1 \\ \exp\left(\frac{g}{\delta} - \bar{u}\right) & \text{if } \gamma = 1 \end{cases} \tag{3}$$

The comparative statics are then straightforward. The higher is the AIBC ratio g/δ, the higher is c^*; faster growth from A.I. raises c^* while a higher rate of existential risk lowers c^*. A higher $\bar{u}$ means that life is more valuable at any given level of consumption, and this reduces c^*; the existential risk is less worth it.

The solution for c^* then implies the optimal choice of T^* since $c = c_0 e^{gT}$:

$$T^* = \frac{1}{g}\log(c^*/c_0).$$

2.2 Quantitative Analysis

The AIBC ratio g/δ is obviously a critical input into any quantitative analysis of this model. We consider each of g and δ in turn. Letting the A.I. run for one additional period raises consumption by, for example, $g = 10\%$. This is extraordinarily rapid economic growth, much faster than the 2% per year growth experienced in the U.S. for the past 150 years. In a semi-endogenous growth setup, achieving this faster growth rate would involve increasing the growth rate of researchers by at least a factor of 5. This would be an amazing accomplishment, but it is one that some observers think is possible for A.I. By choosing such a high value, we are giving the benefit of the doubt to the possibility that A.I. is incredibly useful.

What is the flow probability of existential risk from that action? Experts disagree about this risk in general, but let me consider two possible values to illustrate some important points. First, perhaps the existential risk is 1% per year. Second, perhaps it is twice as dangerous at 2% per year. These values are completely made up, but they are illustrative and the tradeoffs they imply will be clear. The model in the next section takes an alternative approach that sidesteps an assumption like this. In the first case,

Table 1: Consumption and Existential Risk: Simple Model

	$\delta = 1\%$			$\delta = 2\%$		
γ	c^*	T^*	Exist.Risk	c^*	T^*	Exist.Risk
1	54.60	40.0	0.33	1	0	0
2	1.57	4.5	0.04	1	0	0
3	1.27	2.4	0.02	1	0	0

Note: The table shows the quantitative results for the optimal choices from the simple model, assuming $g = 10\%$ so that the AIBC ratio is 10 in the left panel and 5 in the right panel. Other values assumed are $c_0 = 1$ and $v(c_0) = 6$. The value of $\bar{u}$ is chosen to match $v(c_0) = 6$ for each value of γ. The "Exist.Risk" column reports $1 - \exp(-\delta T^*)$, which is the overall probability of existential risk.

the AIBC ratio is 10 while in the second case it is 5. Table 1 shows the quantitative results for various parameter values.

Log utility. As explained earlier, $v(c_{us,today}) = 6$. If $\delta = 1\%$ so that the AIBC ratio is 10, then we would use the A.I. for a number of years until the value of life rises to 10x consumption from its current value of 6x. With log utility ($\gamma = 1$), recall from equation (2) that $v(c) = \bar{u} + \log c$. In this case, $\log c$ would need to increase by 4 units, and $\exp(4) \approx 55$. In other words, with log utility and $\delta = 1\%$, we should run the A.I. until consumption increases by a factor of 55! Growing at 10% per year, this implies $T^* = 40$, so we would grow at this rapid rate for 40 years. By comparison, the United States has experienced an approximately 18-fold increase in GDP per capita since 1870. Modern living standards in the U.S. are also around 50 times higher than those in the least developed nations, which in turn is not much greater than the living conditions experienced by the majority of the world's population throughout most human history.

What is the price of this amazing change in living standards? Recall that we would face a flow probability of existential risk of 1% per year for 40 years, so the probability we survive this A.I. explosion is $\exp(-.01 \times 40) \approx 0.67$. In other words, with log utility it is optimal to take a 1 in 3 chance of ending human existence in exchange for a 2/3 chance of dramatically raising living standards by a factor of 55.

The next interesting finding in Table 1 is what happens with log utility if $\delta = 2\%$

instead of 1%. In this case, notice that the AIBC ratio is 5 instead of 10. But because $v(c_{us,today}) = 6$, the value of life in the U.S. today is already too high to make the A.I. risk worthwhile: $\delta v(c_{us,today}) > g$ so that the optimal choice is $T^* = 0$. Our range of uncertainty about the nature of existential risk surely includes both 1% and 2% for δ. In the former case, we run the A.I. for 40 years and incomes rise by a factor of 55, but in the latter case we optimally shut the A.I. down immediately. This result is summarized in our first key point:

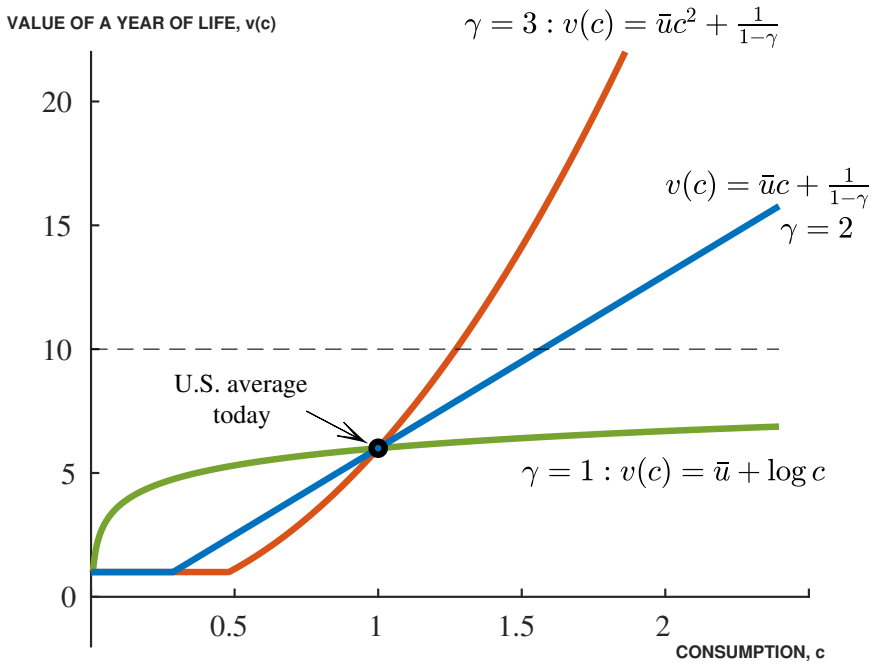
Key Point 1 (Log utility): Decisions and optimal outcomes in the simple model with log utility are very sensitive to the magnitude of the A.I. existential risk. With $\delta = 1\%$ it is optimal to use the A.I. technology for 40 years involving an overall 1/3 probability of existential risk and a stunning 55-fold increase in consumption. With $\delta = 2\%$, it is optimal to shut it down immediately even with log utility (and also for $\gamma > 1$).

CRRA utility with $\gamma > 1$. In the log case, $u(c)$ is not bounded and the value of life rises slowly, with the log of consumption. When $\gamma > 1$, flow utility is bounded and the value of life rises as a power function of consumption — rising linearly with c when $\gamma = 2$; more on this shortly. Our second main point is that the optimal value of T^* is very sensitive to $\gamma = 1$ versus $\gamma = 2$.

To see this, return to the case of $\delta = 1\%$ so that the AIBC ratio is 10 and consider the results in Table 1. It is once again optimal to have $v(c)$ rise from the initial value of 6 to a new value of 10. However, because the value of life rises faster with c, this involves just 4.5 years of A.I.-enabled growth rather than 40 years. Optimal consumption rises by 57% — a factor of 1.57 instead of a factor of 55 — and the economy bears the price of existential risk equal to 4%. Moving to $\gamma = 3$ roughly cuts these values in half. Our second key point summarizes this lesson:

Key Point 2 (CRRA > 1): Holding constant the rate of existential risk at $\delta = 1\%$, the optimal decision varies sharply with whether $\gamma = 1$ (log utility) or $\gamma = 2$. Log utility involves the 55-fold gain in consumption, 40 years of using the A.I., and a 1/3 probability of an existential disaster. With $\gamma = 2$, the gain in consumption is dramatically smaller, 57 percent instead of a factor of 55, the A.I. is used for 4.5 years, and the optimal probability of an existential disaster is just 4 percent.

Figure 2: The Value of a Year of Life, $v(c)$



Note: The value of a year of life is $v(c) \equiv \frac{u(c)}{u'(c)c} = \bar{u}c^{\gamma-1} + \frac{1}{1-\gamma}$. It is the value of living a year $u(c)$, converted into consumption units by dividing by $u'(c)$, expressed as a ratio to c itself. Therefore it has the units of "the value of a year of life measured in years of per capita consumption." In the U.S. today, the value for an average person equals 6; we choose different values of $\bar{u}$ for the different values of γ to match this fact. In the graph, average U.S. consumption today is normalized to 1. As Rosen (1988) pointed out, $v(c)$ cannot be less than one: certainly you'd be willing to give up your consumption... to have your consumption.

2.3 Heterogeneity and the Value of Life

To understand these results better, it is helpful to consider the value of a year of life, $v(c)$. Recall that this value is $v(c) \equiv \frac{u(c)}{u'(c)c} = \bar{u}c^{\gamma-1} + \frac{1}{1-\gamma}$, so it increases with consumption when $\gamma \geq 1$. (In the log case, $v(c) = u(c) = \bar{u} + \log c$.) In the U.S. today, the value for an average person, with consumption around $40,000, equals 6 years of consumption.

Figure 2 plots this value of life against consumption. In this graph, we normalize the units of consumption so that $c_{us,today} = 1$. A key point of the graph is the heterogeneity in the value of life, both as a function of c and as a function of the risk aversion parameter γ.

For example, when $\gamma = 1$, the value of life rises very slowly — with the log of consumption. To get to a $v(c) = 10$ requires a massive increase in c; this was the factor of

55 shown above.

In contrast, when $\gamma = 2$, the value of life rises linearly in consumption. Because $\bar{u} = 7$ in this case, getting $v(c)$ to rise by 4 units from 6 to 10 only requires a 57% increase in c.

Finally, for higher values of γ, $v(c)$ rises even faster: recall that it looks like $v(c) \approx \bar{u}c^{\gamma-1}$. So if $\gamma = 3$, the value of life rises with the square of consumption and if $\gamma = 5$, the value of life rises with c^4.

An implication of this analysis is that people with different levels of consumption or people with different values of γ will feel very differently about using A.I. Consider the low levels of consumption in the poorest countries of the world or for low-income people in the United States. In this case, the marginal utility of consumption is high and these people would be more willing to undertake gambles with their lives in order to reach much higher living standards. On the other hand, people who are rich or people who are very risk averse would be much less willing to take such gambles.

3. Model #2: Mortality Improvements and Singularities

If A.I. doesn't destroy the world, it could do more than accelerate growth in consumption. An A.I. capable of accelerating growth to 10% per year might also create cures for cancer and heart disease and create other innovations that reduce mortality. As we get richer and life becomes more valuable, these mortality reductions could be a key part of the benefits of A.I. The existential risk is partially offset by increasing life expectancy in the good state of the world.

In this section, we incorporate this consideration as well as a more explicit model of dynamics and even the possibility of a singularity in which A.I. leads consumption to go to infinity in finite time.

3.1 The Model

Suppose that utilitarian social welfare is the discounted sum of the flow utilities of everyone alive:

$$U = \int_0^\infty N_0 e^{-(\rho-b+m)t} u(c_t) dt$$

where b is the birth rate at which new people enter the economy, m is the idiosyncratic mortality rate for individuals, and consumption per person grows at rate g: $c_t = c_0 e^{gt}$. The population growth rate is therefore $b-m$, and all three of b, m, and g are exogenous and constant with $\rho - b + m > 0$. Assume CRRA utility, as before, so that $u(c) = \bar{u} + c^{1-\gamma}/(1-\gamma)$ and let us maintain $\gamma > 1$ throughout this section.

Substituting in the utility function and the constant exponential consumption growth, we can solve the integral in the utility function to get

$$U(g,m) = \frac{N_0 \bar{u}}{\rho - b + m} + \frac{N_0 c_0^{1-\gamma}}{1-\gamma} \cdot \frac{1}{\rho - b + m + (\gamma - 1)g}. \tag{4}$$

In the absence of A.I., the economy experiences a constant growth rate given by g_0 and a mortality rate m_0 (the 0 subscripts denote the economy in the absence of A.I., not time subscripts). Adopting the A.I. technology leads to faster economic growth at rate g_{ai} and potentially lower mortality at rate m_{ai}. However, the cost is a one-time existential risk that is realized immediately when the A.I. technology is implemented: with probability δ, every human dies.

In this case, a social planner maximizing expected utility will implement the A.I. as long as

$$U(g_0, m_0) < (1-\delta) U(g_{ai}, m_{ai}).$$

Clearly, then, it is optimal to use the A.I. technology provided the existential risk δ is lower than a critical value δ^* that makes this equation hold with equality:

$$\delta^* = 1 - \frac{U(g_0, m_0)}{U(g_{ai}, m_{ai})}. \tag{5}$$

To summarize, if the actual one-time existential risk from A.I., δ, is smaller than the cutoff δ^*, then it is optimal to use the A.I. On the other hand, if the one-time risk is larger than δ^*, then the A.I. is too dangerous and it is optimal not to use it.

Basic solution. It is helpful to make one additional substitution into equation (4) for $U(g,m)$ before plugging in to solve for δ^*. In particular, recall that $v(c) = \bar{u}c^{\gamma-1} + 1/(1-\gamma)$ so that

$$\bar{u}c^{\gamma-1} = v(c) - \frac{1}{1-\gamma}.$$

This equation can be used to write $U(g,m)c_0^{\gamma-1}$ as a function of $v(c_0)$, which we observe, instead of $\bar{u}$. Substituting into (5) then gives

$$1-\delta^* = \frac{\frac{v(c_0)-\frac{1}{1-\gamma}}{\rho-b+m_0}+\frac{1}{1-\gamma}\cdot\frac{1}{\rho-b+m_0+(\gamma-1)g_0}}{\frac{v(c_0)-\frac{1}{1-\gamma}}{\rho-b+m_{ai}}+\frac{1}{1-\gamma}\cdot\frac{1}{\rho-b+m_{ai}+(\gamma-1)g_{ai}}} \tag{6}$$

Singularity. Next, notice a remarkable fact: when $\gamma > 1$, a singularity that delivers infinite consumption does not deliver infinite utility because flow utility is bounded at $\bar{u}$. This can be seen easily back in our earlier Figure 1. In this case, utility under a singularity that occurs at time 0 is given simply by

$$U_{sing} = \frac{N_0\bar{u}}{\rho-b+m_{ai}},$$

which is also the solution that emerges in (4) when evaluated at $g=\infty$.

Using this logic, we can set $g_{ai}=\infty$ in (6) to solve for the existential risk cutoff when A.I. is associated with a singularity at date 0:

$$1-\delta^*_{sing} = \frac{\rho-b+m_{ai}}{\rho-b+m_0} - \frac{1}{1+(\gamma-1)v(c_0)}\cdot\frac{\rho-b+m_{ai}}{\rho-b+m_0+(\gamma-1)g_0}. \tag{7}$$

Finally, just for intuition, it is helpful to solve for the singularity cutoff when A.I. has no additional mortality benefit so that $m_{ai}=m_0\equiv m$:

$$\delta^*_{sing,m} = \frac{1}{1+(\gamma-1)v(c_0)}\cdot\frac{1}{1+\frac{(\gamma-1)g_0}{\rho-b+m}} \tag{8}$$

The comparative statics are clear from this last equation. A higher initial value of life $v(c_0)$ reduces the existential risk cutoff. A higher normal growth rate g_0 reduces the cutoff. Higher risk aversion γ — sharper diminishing marginal utility — also reduces the cutoff. A higher discount rate ρ or mortality rate m raise the singularity cutoff as the future benefits of regular growth g_0 count for less in outweighing the infinite consumption of the singularity.

3.2 Quantifying the Richer Model

We now quantify the existential risk cutoff δ^* for various cases using the results we've just derived. We start the economy off as before with $v(c_0) = 6$ and assume an effective rate of time preference of $\rho - b = 1\%$.

In the case in which A.I. is not used, we assume $g_0 = 2\%$ and $m_0 = 1\%$, corresponding to consumption growth of 2% per year and a mortality rate of 1% per year, implying a life expectancy of 100 years.

We allow the successful use of A.I. to affect growth in one of two ways: a fast-growth scenario with $g_{ai} = 10\%$ or an immediate singularity that delivers infinite consumption. With respect to mortality, we also consider two scenarios. In the first, A.I. does not affect mortality and $m_{ai} = m_0 = 1\%$. In the second, we assume that the innovative A.I. capable of astounding consumption growth also can offer impressive mortality improvements so that the mortality rate falls in half to $m_{ai} = 0.5\%$. Notice that this corresponds to life expectancy doubling to 200 years, so this is a large change.

The results for the existential risk cutoff δ^* are shown in Table 2. The first row considers $\gamma = 1.01$, very close to log utility. This row confirms the results we saw in the simple model: with log utility, optimal existential risk cutoffs are remarkably high. For example, when the A.I. delivers 10% growth and no mortality improvement, we find $\delta^* = 35\%$. Also paralleling the simple model, as we increase γ to 2 or 3, the existential risk cutoff falls very sharply to just 4.9% and 1.9% respectively. So the first column of Table 2 basically confirms the results we saw earlier.

Singularities. Next, consider the consequences of making A.I. even more impressive, so that it leads to an immediate singularity with infinite consumption. These results are shown in the right panel of Table 2. With near-log utility, the existential risk cutoff for implementation gets very large, approaching 100%. In fact, it is easy to show that with $\gamma \leq 1$ — so that utility is logarithmic or even less curved — the optimal existential risk cutoff for a singularity is 100%. That is, as long as total annihilation of the human race is not a sure thing, the infinite consumption dominates and A.I. implementation maximizes this social welfare function. This strikes me as implausible, which is consistent with a large literature in economics focusing on $\gamma > 1$ instead of $\gamma \leq 1$.

The middle and bottom row of the right panel of the table show that $\gamma = 2$ com-

Table 2: Existential Risk Cutoffs: Mortality Improvements and Singularities

	Fast growth: $g_{ai} = 10\%$		Singularity: $g_{ai} = \infty$	
	— m_{ai} —		— m_{ai} —	
γ	1%	0.5%	1%	0.5%
1.01	0.350	0.572	0.934	0.951
2	0.049	0.290	0.071	0.304
3	0.019	0.265	0.026	0.269

Note: The table shows the quantitative results for the existential risk cutoff δ^* in the model with mortality improvements and singularities using equation (6). In the absence of A.I. use, we assume $g_0 = 2\%$ and $m_0 = 1\%$. Other assumed parameter values are $\rho - b = 1\%$ and $v(c_0) = 6$.

pletely changes the story. Because flow utility is bounded, infinite consumption is not that much better than $g_{ai} = 10\%$, and δ^* falls to around 7%. With $\gamma = 3$, the decline is even sharper to $\delta^* = 2.6\%$. These findings lead to our third key point:

Key Point 3 (Singularities): How much existential risk society is willing to bear depends critically on whether or not flow utility is bounded. If $\gamma \leq 1$, the existential risk cutoff for an immediate singularity that delivers infinite consumption is $\delta^* = 1$: any risk other than sure annihilation is acceptable to achieve infinite consumption. In contrast, if $\gamma = 2$ or 3, the singularity cutoffs are much closer to the cutoffs with $g_{ai} = 10\%$ and are much smaller. For example $\delta^* = 2.6\%$ when $\gamma = 3$: even infinite consumption is not worth the gamble if the one-time existential risk is greater than 2.6%.

Improved mortality. Finally, consider the possibility of mortality improvements. The innovations that A.I. creates to accelerate economic growth may affect more than just consumption. We already see examples of A.I. being used for protein folding, drug discovery, and evaluating images. An A.I. that could accelerate consumption growth to 10% or more would surely create innovations that also reduce mortality. As we get richer and life becomes more valuable, these mortality reductions could be a key part of how A.I. improves living standards. The existential risk may be partially balanced by letting everyone live longer in the good state of the world where the existential risk is

not realized.

Table 2 illustrates the importance of this force by considering the possibility that A.I. cuts the standard mortality rate in half, from 1% per year to 0.5% per year. The effects on the optimal cutoff for existential risk, δ^*, are large. The intuition for this is that mortality reductions are "in the same units" as existential risk, unlike consumption which gets filtered through a bounded utility function. When $\gamma = 2$ and $g_{ai} = 10\%$ for example, the cutoff for using A.I. (δ^*) rises sharply from 4.9% to 29.0%. When $\gamma = 3$, the change is even more dramatic, with δ^* rising from 1.9% to 26.5%. A 1-in-4 chance of an existential catastrophe is more bearable when we live for 200 years instead of 100 years if the catastrophe does not occur.

The point that mortality and existential risk are in the same units can be made even more clearly in the simple model from Section 2. For example, when the existential risk has arrival rate δ, the probability of surviving T years is $S(T) = \exp[-(\delta + m)T]$ and it is only the sum of existential and mortality risk that matters, not the composition.

The insights regarding mortality improvements are summarized in our fourth key point:

> *Key Point 4 (Mortality improvements): With $\gamma > 1$, consumption gains have sharply diminishing returns and life becomes increasingly valuable. If A.I. can also improve life expectancy and mortality rates apart from existential risk, the existential risk cutoffs are much higher, on the order of 25–30% for $\gamma = 2$ or 3.*

Longtermism. What happens if we discount the future at a lower rate and therefore put more weight on the future? Two insights emerge, one relatively obvious and one less so.

First, return to the case where $m_{ai} = m_0$ so A.I. does not generate any mortality improvements. Consider what happens if the effective discount rate $\rho - b + m$ falls to zero. This is easiest to see in equation (8). If $\rho - b + m \to 0$, then $\delta^*_{sing,m} \to 0$. That is, if we are risking an infinite future that is effectively undiscounted, the existential risk cutoff falls to zero. It is not worth any one-time existential risk even to achieve a singularity because an infinity of futures is at risk (and because the singularity itself only delivers finite utility). This echoes the "What We Owe the Future" effect of MacAskill (2022).

Second, consider what happens to the value of mortality improvements as we place

Table 3: Existential Risk Cutoffs: Mortality Improvements with Less Discounting

	Baseline $\rho - b = 1\%$		Less discounting $\rho - b = -0.45\%$	
	— m_{ai} —		— m_{ai} —	
γ	1%	0.5%	1%	0.5%
1.01	0.934	0.951	0.910	0.992
2	0.071	0.304	0.031	0.912
3	0.026	0.269	0.009	0.910

Note: The table shows the quantitative results for the existential risk cutoff δ^* in the model with singularities using equation (7). We assume $g_0 = 2\%$, $g_{ai} = \infty$, $m_0 = 1\%$, and $v(c_0) = 6$.

more weight on the future. As before, assume A.I. lowers the mortality rate from 1% to 0.5%. In this case, we have two different effective discount rates, $\rho - b + m_0$ and $\rho - b + m_{ai}$. To keep everthing finite, suppose we lower $\rho - b$ to -0.45%. The effective discount rate with A.I. becomes nearly zero at $\rho - b + m_{ai} = 0.05\%$ while the rate in the absence of A.I. is a half percentage point higher at $\rho - b + m_0 = 0.55\%$. The results are shown for the singularity case in Table 3. (The results for $g_{ai} = 10\%$ are very similar.)

The first two columns repeat our baseline calculation with $\rho - b = 1\%$ for easy comparison. The third column considers lowering $\rho - b$ to -0.45% holding constant $m_{ai} = m_0 = 1\%$. As expected, the prize to be lost from existential risk is larger, so the cutoffs decline.

The surprise comes in the last column, where we study the case in which A.I. improves the mortality rate so that $m_{ai} = 0.5\%$. The cutoffs rise sharply to $\delta^* > 0.9$. Any one-time existential risk of less than 90% is worth taking in order to improve the mortality rate by half a percentage point, even for γ as high as 3. The reason is that we are discounting the future less, so mortality improvements themselves are more valuable.[1]

The key intuition is that as $\rho - b + m_{ai} \to 0$, the social welfare from adopting A.I., $U(g_{ai}, m_{ai})$, goes to infinity, while the welfare from not adopting A.I., $U(g_0, m_0)$ stays

[1]To see this mathematically, return to equation (7) and factor out $\rho - b + m_{ai}$. The limiting results are the same with finite g_{ai}, but the derivatives are not always monotonic.

finite — the higher mortality rate m_0 ensures a postive effective discount rate in that case. With a common mortality rate, both discount rates fall to zero and both social welfares go to infinity. But if A.I. reduces the mortality rate, it is the A.I. case that delivers infinite welfare and hence makes any existential risk other than sure disaster worth taking.

Even with a longtermism focus, society in this example is willing to tolerate huge existential risk if one benefit of A.I. is to reduce mortality and improve life expectancy. We value the large number of future generations, and they themselves benefit tremendously from living longer. Extending lives from 100 to 200 years is not especially valuable if we heavily discount the future, but it becomes tremendously valuable with a longterm focus.

Key Point 5 (Longtermism): Consider the case where A.I. leads to a singularity. Absent mortality improvements, lowering the effective discount rate to place more weight on the future reduces the existential risk cutoff, which falls to zero in the limit. With mortality improvements, the result is the opposite: putting more weight on the future means that A.I.-driven mortality improvements are more valuable, ultimately making any existential risk other than sure disaster worth bearing.

4. Conclusion

The point of this paper is not to provide a sharp answer to the question of "Should we shut down A.I.?" even setting aside the important issue of how that could be achieved. Instead, simple models are used to study how the answer to this question varies with relatively small changes in how we set up the problem.

One key sensitivity is whether we use log utility or CRRA utility with $\gamma = 2$ or more. With log utility, remarkably large amounts of existential risk are tolerated in order to take advantage of huge advances in living standards. But with $\gamma = 2$ or more, gambling with existential risk is much less appealing.

Next, even singularities that deliver infinite consumption immediately are not as valuable as one might have thought. With bounded utility (e.g. $\gamma > 1$), infinite consumption merely pushes us to the upper bound and the marginal utility of the additional consumption is small. The finding that with $\gamma = 2$ or more, social welfare in

these models suggests taking great care with existential risk continues to hold even in the presence of a singularity.

Finally, one way in which it can be optimal to entertain greater amounts of existential risk is if A.I. leads to new innovations that improve life expectancy. Mortality improvements and existential risk are measured in the same units and do not run into the diminishing marginal utility of consumption. This result is reinforced by low effective discount rates that put high weight on the future.

There are of course many considerations that are omitted from this analysis. For example, investments in A.I. safety may lower existential risk. It may be optimal to delay using the A.I. until the risk can be lowered (even beyond the cutoffs here, depending on the effectiveness of those investments). Another consideration involves the nature of risk. Here, individuals — and hence the utilitarian social planner — treat "10% of the population dies each period" as equivalent to "there is a 10% chance of human extinction" because from an individual's standpoint, both involve a 10% chance of dying. In contrast, many people have the instinct that these two risks should not be symmetric, which could lead to more conservative cutoffs.

References

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," in Ajay Agrawal, Joshua Gans, and Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*, University of Chicago Press, 2019, pp. 237–282.

Aschenbrenner, Leopold, "Existential Risk and Growth," September 2020. Global Priorities Institute Working Paper No. 6-2020.

Bostrom, Nick, "Existential Risks," *Journal of Evolution and Technology*, March 2002, *9*.

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

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," March 2023.

Davidson, Tom, "Could Advanced AI Drive Explosive Economic Growth?," June 2021. Open Philanthropy report.

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

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

Joy, Bill, "Why the Future Doesn't Need Us," *Wired Magazine*, April 2000, *8* (4).

MacAskill, William, *What We Owe the Future*, Basic books, 2022.

Martin, Ian W. R. and Robert S. Pindyck, "Welfare Costs of Catastrophes: Lost Consumption and Lost Lives," *The Economic Journal*, 08 2020, *131* (634), 946–969.

Martin, Ian W.R. and Robert S. Pindyck, "Averting Catastrophes: The Strange Economics of Scylla and Charybdis," *American Economic Review*, October 2015, *105* (10), 2947–85.

Murphy, Kevin M. and Robert Topel, "The Economic Value of Medical Research." In *Measuring the Gains from Medical Research: An Economic Approach* Murphy and Topel, eds (2003) pp. 41–73.

__ and __ , eds, *Measuring the Gains from Medical Research: An Economic Approach*, Chicago: University of Chicago Press, 2003.

Ngo, Richard, Lawrence Chan, and Sören Mindermann, "The Alignment Problem from a Deep Learning Perspective," 2023.

Nordhaus, William D., "The Health of Nations: The Contribution of Improved Health to Living Standards." In Murphy and Topel, eds (2003) pp. 9–40.

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

Rees, Martin, *Our Final Century*, London: William Heinemann, 2003.

Rosen, Sherwin, "The Value of Changes in Life Expectancy," *Journal of Risk and Uncertainty*, 1988, *1*, 285–304.

Trammell, Philip and Anton Korinek, "Economic Growth under Transformative AI," 2020. GPI Working Paper No. 8-2020.

Yudkowsky, Eliezer et al., "Artificial intelligence as a positive and negative factor in global risk," *Global catastrophic risks*, 2008, *1* (303), 184.