The A.I. Dilemma: Growth versus Existential Risk

Charles I. Jones*
Stanford GSB and NBER

April 17, 2024 — Version 2.0

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. On the other hand, many experts worry that these advances entail existential risk: creating a superintelligence misaligned with human values could lead to catastrophic outcomes, even possibly 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, Maya Eden, Joshua Gans, Tom Houlden, Pete Klenow, Anton Korinek, Kevin Kuruc, Pascual Restrepo, Charlotte Siegmann, Chris Tonetti, Phil Trammell and seminar participants at the Markus Academy, the Minneapolis Fed, the NBER A.I. conference, Oxford, PSE Macro Days 2023, Stanford, and USC for helpful comments and discussions.

1. Introduction

Recent advances in artificial intelligence (A.I.) will likely 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 match or even exceed human intelligence at many cognitive tasks and 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 so-called "singularity" with infinite consumption. Models along these lines have been explored by Aghion, Jones and Jones (2019), Trammell and Korinek (2020), Davidson (2021), Nordhaus (2021), and Erdil and Besiroglu (2023).

On the other hand, these advances do not come without risk. A substantial contingent of the A.I. community, including leading researchers at OpenAI and Google, warn that these advances could constitute an existential risk for humanity, either from malicious use of the A.I. by a "bad actor" or perhaps even from a superintelligent A.I. itself.

More succinctly, A.I. could raise living standards by more than electricity or the internet. But it may pose risks that exceed those from nuclear weapons. Moreover, these possibilities — however likely or unlikely — are correlated. It is precisely the state of the world in which A.I. is sufficiently powerful to generate profound increases in living standards that seems most likely to pose existential risk. 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.

Several insights emerge:

1. The curvature of utility is very important. With log utility, the models are remarkably unconcerned with existential risk, suggesting that the 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. Such models are quite conservative in trading off consumption gains versus existential risk.

3. These findings extend to singularity scenarios in which self-improving A.I. generates accelerating growth and infinite consumption. 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 finite utility. The models with $\gamma \geq 2$ remain conservative with regard to existential risk.

4. A key exception emerges if the rapid innovation facilitated by A.I. extends life expectancy. These gains are "in the same units" as existential risk in the sense that they 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.

Section 2 develops a simple model to illustrate these forces. Section 3 then extends the analysis to include a richer theory of dynamics, the possibility of a singularity, and the prospect that A.I. innovations extend life expectancy.

Related literature. Serious concerns about the existential risk associated with A.I. have been highlighted by Joy (2000), Bostrom (2002, 2014), Rees (2003), Posner (2004), Yudkowsky et al. (2008), and Ord (2020). These concerns have accelerated together with the progress of A.I. itself. Ngo, Chan and Mindermann (2023) provides a recent overview.

A recent conference volume on the economics of A.I. (Agrawal, Gans and Goldfarb, 2019) highlights a range of potential costs and benefits. Brynjolfsson and McAfee (2014) emphasize large benefits from A.I., while Brynjolfsson, Rock and Syverson (2021) note that growth could initially slow as organizational changes are implemented, reminiscent of the adoption of electricity and information technology. Acemoglu and Lensman

(2023) show that the optimal adoption of transformative technologies that involve large costs and benefits can be delayed if the costs are somewhat irreversible.

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 risks associated with frontier science. As we get richer, it may be optimal to slow economic growth, or at least redirect innovation toward life-saving technologies. The present paper differs by focusing explicitly on A.I. and in quantifying the tradeoffs.

Aschenbrenner (2020) and Aschenbrenner and Trammell (2024) focus on existential risk more generally, positing that it increases with aggregate consumption and decreases with mitigation efforts. They suggest we may live in a "time of perils" in which we are advanced enough to face high risk but not rich enough to spend sufficiently on mitigation. Martin and Pindyck (2015, 2020) consider catastrophes and how the value of a statistical life 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

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, perhaps 10% per year. However, the use of this A.I. poses an existential risk to humanity. Using the advanced A.I. for T periods raises consumption per person to $c = c_0 e^{gT}$, but at the same time, the probability that humanity 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. (though we will sometimes loosely refer to this as "how many years"). All growth and existential risk is realized immediately rather than over time, and if society survives, people consume the constant $c = c_0 e^{gT}$ 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 social welfare up or down 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 $1/\rho$.

From the first order condition, it is optimal to use the A.I. as long as

$$\underbrace{\delta \cdot \frac{N}{\rho} v(c)}_{\text{Lost lives}} \leq \underbrace{g \cdot \frac{N}{\rho}}_{\text{Extra growth}} \quad \text{where } v(c) \equiv \frac{u(c)}{u'(c)c}.$$

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 $u(c)$ per person, scaled up by N/ρ for population and present value; we divide by $u'(c)$ to convert utils to consumption units and consider the ratio to c because the tradeoff is with a higher growth rate, which itself is a percent of consumption. So the value of a lost period of life is $v(c) \equiv \frac{u(c)}{u'(c)c}$. The benefit of using the A.I. more intensively is the extra period of consumption growth at rate g, also scaled by N/ρ. The optimal choice of how intensively to use the A.I. trades off these costs and benefits.

Canceling terms, rewriting, and assuming an interior solution, the optimal choice is T^* such that $c^* = c_0 e^{gT^*}$ satisfies

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

The left side of this equation, $v(c) \equiv u(c)/u'(c)c$, is the value of a year of life in consumption units as a ratio to consumption per person. For example, in the United States today, a typical value of a life year is around $250,000, which comes from a "value of a statistical life (VSL)" 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 (Hall,

Jones and Klenow, 2020). The precise value could be 25% higher or lower depending on assumptions; nothing critical hinges on the exact value.

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, 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

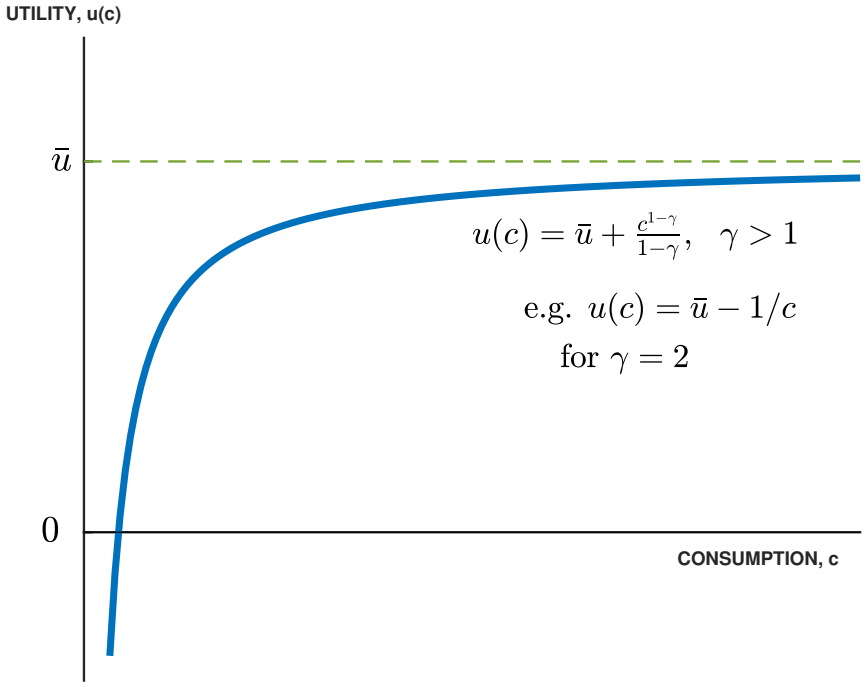
$$v(c) \equiv \frac{u(c)}{u'(c)c} = \begin{cases} \bar{u}c^{\gamma-1} + \frac{1}{1-\gamma} & \text{if } \gamma \neq 1 \\ \bar{u} + \log c & \text{if } \gamma = 1 \end{cases} \tag{2}$$

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 (Kimball et al., 2024). 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, 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.

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 $\bar{u}$.

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 (Jones, 2022). This would be an amazing accomplishment, but it is one that some observers think is possible for A.I. (Davidson, 2021). 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 the need to calibrate this parameter. In the first case, 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 poorest countries, which in turn are not much greater than living standards experienced by most people in the world before the Industrial Revolution.

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

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.

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%. In the former case, we run the A.I. for the equivalent of 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): Results 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. for 40 years involving a 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.

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.

Key Point 2 (CRRA > 1): With $\gamma = 2$, it remains 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 optimal probability of an existential disaster is just 4%. Moving to $\gamma = 3$ roughly cuts these values in half.

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 is about 6x annual 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.

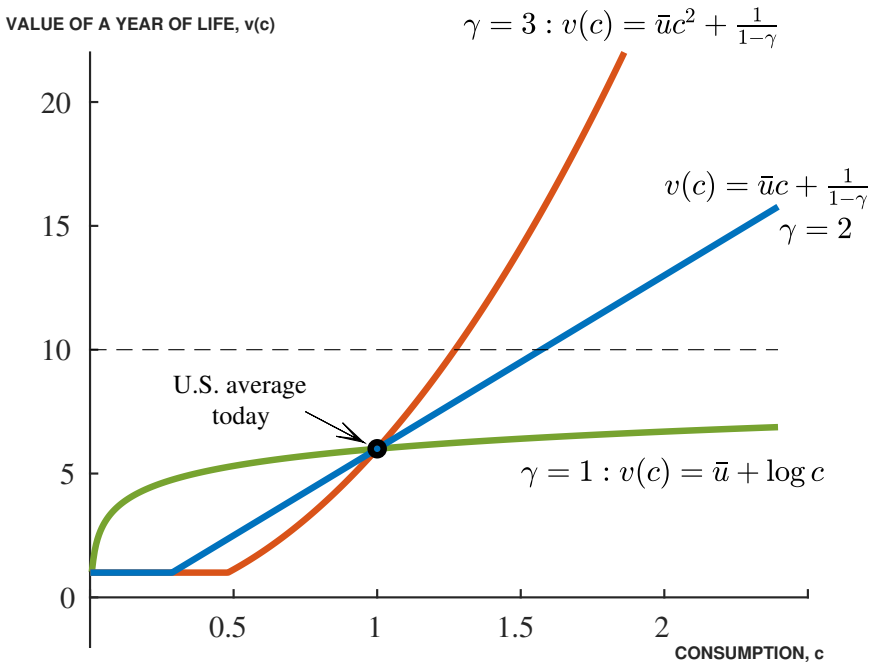
For example, when $\gamma = 1$, the value of life rises very slowly — with the log of consumption. To get to a $v(c) = 10$ requires a massive increase in c; this was the factor of 55 mentioned in Key Point 1.

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 lower levels of consumption in China or in the poorest countries of the world, or for low-income people in the United States. In this case, the marginal utility of

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 as a multiple of individual 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, there are reasons why $v(c)$ may not be less than one: certainly you'd be willing to give up your consumption... to have your consumption.

consumption is high and these people would be more willing to undertake gambles with their lives in order to reach much higher living standards. The same would be true for people with low degrees of risk aversion, perhaps including entrepreneurs at the very tech startups developing A.I. On the other hand, people who are rich or people who are very risk averse would be much less willing to take such gambles.

2.4 Discussion

Now is a good time to pause and discuss the assumptions behind the approach taken here. For example, consider the question of whether a marginal VSL can be used to tell us something about existential risk. Recall that the VSL is typically measured using compensating differentials in the labor market: how much higher a wage does a worker demand for being a construction worker, facing a high mortality risk, than for being a janitor? This tells us how the individual values the additional 40 years of $u(c)$ flows, and we can use this to get the "value of a statistical life year" corresponding to the value of a single year of $u(c)$. In our simple model, that is what we need. A constant population of N people each enjoy a constant $u(c)$ each period of life. All individuals face the same tradeoff between existential risk and higher consumption. Since all people are identical, the way in which a representative person trades off mortality risk and higher consumption is exactly informative about this tradeoff.

One way in which the VSL might overstate the value of existential risk is if construction workers have families who might suffer if the construction worker dies. This is already included in VSLs, but with existential risk no one remains to suffer. This argument suggests using lower value-of-life numbers to value existential risk, implying that the risk is more worth bearing.

One might also object that ending the existence of all human life on earth cannot be properly measured by extrapolating from the value of ending a single person's life. The approach we've implicitly taken so far and will make more explicit in the next section is a total utilitarian perspective that simply adds up across all the lifetime utilities of people who might live. This adding up implies a certain linearity that is worth discussing.

First, notice that this total utilitarian perspective captures exactly the logic that trillions of future people might not get to live, and that this loss would be a tragedy. It values the trillions of people precisely as they would value living their lives. Is there

something *beyond* this that should be included as well? Perhaps. But the point is that the baseline social welfare function that economists and philosophers often rely on is this total utilitarian perspective; it is valuable to see what it says.

Second, it is worth noting that there are excellent reasons for the simple "adding up" of lifetime utilities to get social welfare. In fact, there are theorems that justify this linearity under weak and seemingly plausible conditions; see Kuruc, Budolfson and Spears (2022). Philosophers therefore consider the total utilitarian perspective to be focal and worthy of attention (Zuber et al., 2021).

3. Model #2: Singularities and Mortality Improvements

We now consider a richer model that doesn't require specifying exactly the probability of existential risk. Interesting results emerge when we consider (1) a singularity in which A.I. quickly leads to infinite consumption, (2) the possibility that A.I. might create cures for diseases and reduce mortality more generally, and (3) near-zero social discounting to put more weight on future generations.

3.1 The Model

Consider utilitarian social welfare — the discounted sum of lifetime utilities:

$$W = \int_0^\infty e^{-\rho_s t} N_t U_t dt \tag{3}$$

where ρ_s is the social discount rate across generations, $N_t = N_0 e^{bt}$ is the number of people born at date t (which grows exogenously at rate b for "births"), and U_t is the expected lifetime utility of an individual from the cohort born at date t:

$$U_t = \int_0^\infty e^{-(\rho+m)a} u(c_{t+a}) da.$$

The parameter ρ is the rate at which individuals discount their own future utility flows. Each individual faces a constant instantaneous probability m of idiosyncratic mortality as in Blanchard (1985) and therefore survives to age a with probability e^{-ma}. Finally, consumption per person is identical across people at a given point in time and grows at rate g: $c_{t+a} = c_0 e^{g(t+a)} = c_t e^{ga}$. All three of b, m, and g are exogenous and constant

with $\rho_s - b > 0$. 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 for expected lifetime utility:

$$U_t = \frac{\bar{u}}{\rho + m} + \frac{c_t^{1-\gamma}}{1-\gamma} \cdot \frac{1}{\rho + m + (\gamma - 1)g}. \tag{4}$$

Finally, we can substitute this expression into the social welfare function in (3) to get

$$W(g, m) = \frac{N_0 \bar{u}}{(\rho + m)(\rho_s - b)} + \frac{N_0 c_0^{1-\gamma}}{1-\gamma} \cdot \frac{1}{(\rho + m + (\gamma - 1)g)(\rho_s - b + (\gamma - 1)g)}. \tag{5}$$

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 A.I. 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.

A social planner maximizing expected social welfare implements the A.I. as long as

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

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

$$\delta^* = \frac{W(g_{ai}, m_{ai}) - W(g_0, m_0)}{W(g_{ai}, m_{ai})}. \tag{6}$$

That is, the existential risk cutoff δ^* is equal to the percentage difference in social welfare between the two situations.

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. If the one-time risk is larger than δ^*, then the A.I. is too dangerous and it is optimal not to use it. A closed form solution for δ^* can be obtained by making the relevant substitutions into (6), but the formula is messy; we will study the general outcomes numerically and the analytic solution for a special case.[1]

[1]It is helpful to make one additional substitution into equation (4) for $W(g, m)$ before plugging in to

Singularity. 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. Social welfare with infinite consumption at time 0 is

$$W_{sing} = \frac{N_0 \bar{u}}{(\rho + m_{ai})(\rho_s - b)}.$$

For intuition, it is helpful to solve for the singularity cutoff when A.I. has no additional mortality benefit so that $m_{ai} = m_0 \equiv m$. In this case, plugging in the expressions for social welfare into (6) gives

$$\delta^*_{sing,m} = \frac{1}{1 + (\gamma - 1)v(c_0)} \cdot \frac{1}{(1 + \frac{(\gamma-1)g_0}{\rho+m})(1 + \frac{(\gamma-1)g_0}{\rho_s - b})} \tag{7}$$

The comparative statics are then clear. A higher initial value of life $v(c_0)$ reduces the existential risk cutoff. A higher starting growth rate g_0 reduces the cutoff. A higher γ — sharper diminishing marginal utility — also reduces the cutoff. On the other hand, a higher social discount rate ρ_s, private discount rate ρ, or mortality rate m raises the singularity cutoff as the future benefits of regular growth g_0 count for less in outweighing the infinite consumption of the singularity. If $g_0 = 0$, the expression simplifies further.

3.2 Quantifying the Richer Model

We now quantify the existential risk cutoff δ^* for various cases. We start the economy off as before with $v(c_0) = 6$ and assume $\rho = 1\%$ and $\rho_s - b = 1\%$: both individuals and the social planner discount the future at a rate of 1% per year (e.g. $b = 0$).

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 a singularity that delivers infinite consumption immediately. 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 delivers impressive mortality im-

solve for δ^*. In particular, recall that $v(c) = \bar{u}c^{\gamma-1} + 1/(1-\gamma)$. This expression can be solved for $\bar{u}$ and used to write $W(g,m)$ as a function of $v(c_0)$, which we observe, instead of $\bar{u}$.

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	0.540	0.678	0.916	0.937
2	0.022	0.267	0.024	0.268
3	0.005	0.254	0.005	0.254

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 $\rho = \rho_s - b = 1\%$ and $v(c_0) = 6$.

provements: 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 (paralleling the large change in growth).

The results for the existential risk cutoff δ^* are shown in Table 2. The first entry considers $\gamma = 1.01$, very close to log utility. This case 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 $\delta^* = 54\%$. Also paralleling the simple model, as we increase γ to 2 or 3, the existential risk cutoff falls very sharply to just 2.2% and 0.5% respectively. So the first column of Table 2 basically confirms the results we saw in the simple model.

Singularities. Next, suppose A.I. is even more impressive, leading 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 rises substantially, 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. adoption maximizes social welfare. This strikes me as unappealing, 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 show that $\gamma = 2$ completely changes the story. Because flow utility is bounded, infinite consumption is not much better than $g_{ai} = 10\%$, and δ^* falls to around 2%. With $\gamma = 3$, the decline is even sharper to $\delta^* = 0.5\%$. These findings lead to our third key point:

Key Point 3 (Singularities): 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 more, the singularity cutoffs are much closer to the cutoffs with $g_{ai} = 10\%$ and are much smaller. For example, when $\gamma = 2$ even infinite consumption is not worth the gamble if the one-time existential risk is greater than 2.4%.

Improved mortality. Next, 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 event that A.I. doesn't cause an existential catastrophe.

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 must be filtered through a bounded utility function. When $\gamma = 2$ and $g_{ai} = 10\%$ for example, the cutoff for using A.I. rises sharply from 2.2% to 26.7%. When $\gamma = 3$, the change is even more dramatic, with δ^* rising from 0.5% to 25.4%. 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 more clearly by drawing on the simple model from Section 2. For example, if the existential risk is a Poisson process with arrival rate δ, the probability an individual survives T

years is $S(T) = \exp[-(\delta + m)T]$; it is only the sum of the risks that matters, not the composition.

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

Near zero social discounting. What happens if the social planner discounts the future at a lower rate, putting 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 and set $b = 0$ so the entering size of each cohort is constant. Consider what happens if the social discount rate ρ_s falls to zero. This is easiest to see in equation (7). If $\rho_s \to 0$, then $\delta^*_{sing,m} \to 0$ (recall that $b = 0$). That is, if we are risking an infinite future that is 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 rich futures is at risk (and because the singularity itself only delivers finite utility with $\gamma > 1$). This echoes the logic of Ord (2020) and MacAskill (2022).

These effects are shown numerically in Table 3. The first two columns repeat our baseline calculation with $\rho_s = 1\%$ and $b = 0$ for easy comparison. The third column considers lowering ρ_s by a factor of 20 to 0.05% but with no mortality improvements. As expected from the discussion just given, the prize to be lost from existential risk is much larger, so the cutoffs decline sharply, almost to zero.

Now consider what happens if A.I. also leads to mortality improvements. As before, assume A.I. lowers the mortality rate from 1% to 0.5%. The last column of Table 3 shows that the cutoffs remain high, around $\delta^* \approx 25\%$ for $\gamma = 2$ or 3 even as ρ_s falls almost to zero. On the one hand, we are putting more weight on future generations. On the other hand, those future generations themselves all value the mortality improvements, and we are discounting the mortality gains for future generations at a lower rate.

The easiest way to understand this result is to recognize that with $\rho_s \to 0$, there are an infinite number of arbitrarily rich future generations who count equally. After all, even with 2% economic growth, future generations will eventually be richer than Bill

Table 3: Existential Risk Cutoffs with Near Zero Social Discounting

	Baseline $\rho_s = 1\%$		Near zero social discounting $\rho_s = 0.05\%$	
	— m_{ai} —		— m_{ai} —	
γ	1%	0.5%	1%	0.5%
1.01	0.540	0.678	0.525	0.646
2	0.022	0.267	0.002	0.251
3	0.005	0.254	0.000	0.250

Note: The table shows the quantitative results for the existential risk cutoff δ^* in the model with singularities using equation (6). We assume $g_0 = 2\%$, $g_{ai} = 10\%$, $m_0 = 1\%$, $b = 0$, and $v(c_0) = 6$.

Gates. In that limit, the utility of each generation is identical and given by

$$U(LE) = \frac{\bar{u}}{\rho + m} = \frac{LE \cdot \bar{u}}{\rho \cdot LE + 1}$$

where $LE \equiv 1/m$ is life expectancy. Notice that if $\rho = 0$, lifetime utility is linear in life expectancy, but with private discounting, lifetime utility features diminishing returns to life expectancy: adding years of life in distant old age is less valuable.

As the social discount rate $\rho_s \to 0$, since the infinity of future generations are the same, we can just think about a "representative" generation. The existential risk cutoff when considering an AI that doubles life expectancy solves

$$U(LE) = (1 - \delta^*)\, U(2 \cdot LE)$$

which implies

$$\delta^* = \frac{1/2}{\rho \cdot LE + 1}.$$

Finally, with $\rho = 0.01$ and $LE = 100$, the solution is $\delta^* = 1/4$, helping to justify the frequent appearance of similar numbers in Table 3. If lifetime utility were linear in life expectancy ($\rho = 0$), then $\delta^* = 1/2$. Alternatively if $\rho \cdot LE = 2$, the cutoff is $\delta^* =$

$1/6$. Importantly, even with $\rho_s \to 0$, these cutoffs are high and far from zero. Even with a future-oriented focus, society in these examples is willing to tolerate substantial existential risk if one benefit of A.I. is to reduce mortality and improve life expectancy.

Key Point 5 (Near zero social discounting.): Absent mortality improvements, lowering the social discount rate to place more weight on the future reduces the existential risk cutoff, which falls to zero in the limit. With mortality improvements, this is no longer the case: putting more weight on future generations means that A.I.-driven mortality improvements are more valuable, potentially making large existential risks 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 depending on 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. (Do the tech entrepreneurs developing A.I. have low risk aversion?) 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, loosely, in the same units and do not run into the diminishing marginal utility of consumption. This result remains true even with low social discount rates that put high weight on future generations.

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; at some point, it may not be possible to lower the risk further, and then calculations like those in this paper apply. Another consideration involves the nature of risk. Here, a utilitarian social planner treats "10% of the population dies each period" as equivalent to "there is a 10% chance of human extinction" because the planner only counts total utils. In contrast, many people have the instinct that these two risks are not equivalent, which could lead to more conservative cutoffs.

References

Acemoglu, Daron and Todd Lensman, "Regulating Transformative Technologies," July 2023. NBER Working Paper 31461.

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.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, University of Chicago Press, 2019.

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

__ and Philip Trammell, "Existential Risk and Growth," February 2024. Global Priorities Institute at Oxford, manuscript.

Blanchard, Olivier J., "Debts, Deficits, and Finite Horizons," *Journal of Political Economy*, 1985, *93* (2), 223–247.

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

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

Brynjolfsson, Erik and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W.W. Norton, 2014.

__ , Daniel Rock, and Chad Syverson, "The Productivity J-Curve: How Intangibles Complement General Purpose Technologies," *American Economic Journal: Macroeconomics*, January 2021, *13* (1), 333–72.

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.

Erdil, Ege and Tamay Besiroglu, "Explosive growth from AI automation: A review of the arguments," 2023.

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.

__ , __ , and Peter J. Klenow, "Trading Off Consumption and Covid-19 Deaths," *Quarterly Review*, 2020, *42* (1).

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

__ , "The Past and Future of Economic Growth: A Semi-Endogenous Perspective," *Annual Review of Economics*, August 2022, *14*, 125–152.

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

Kimball, Miles S., Daniel Reck, Fudong Zhang, Fumio Ohtake, and Yoshiro Tsutsui, "Diminishing Marginal Utility Revisited," Working Paper 32077, National Bureau of Economic Research January 2024.

Kuruc, Kevin, Mark Budolfson, and Dean Spears, "Population Issues in Welfare Economics, Ethics, and Policy Evaluation," 02 2022.

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.

_ , "Are We Approaching an Economic Singularity? Information Technology and the Future of Economic Growth," *American Economic Journal: Macroeconomics*, January 2021, *13* (1), 299–332.

Ord, Toby, *The Precipice: Existential Risk and the Future of Humanity*, Hachette Books, 2020.

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.

Zuber, Stéphane, Nikhil Venkatesh, Torbjörn Tännsjö, Christian Tarsney, H Orri Stefánsson, Katie Steele, Dean Spears, Jeff Sebo, Marcus Pivato, Toby Ord et al., "What should we agree on about the repugnant conclusion?," *Utilitas*, 2021, *33* (4), 379–383.