NBER WORKING PAPER SERIES

HOW MUCH SHOULD WE SPEND TO REDUCE A.I.'S EXISTENTIAL RISK?

Charles I. Jones

Working Paper 33602
http://www.nber.org/papers/w33602

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2025

I am grateful to Daron Acemoglu, David Autor, Maya Eden, Tom Houlden, Matt Kahn, Pete Klenow, Anton Korinek, Toby Ord, Chris Tonetti, Phil Trammell, and to participants at the Economics of Transformative AI conference for helpful feedback and suggestions. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

How Much Should We Spend to Reduce A.I.'s Existential Risk?
Charles I. Jones
NBER Working Paper No. 33602
March 2025
JEL No. O40

ABSTRACT

During the Covid-19 pandemic, the United States effectively "spent" about 4 percent of GDP — via reduced economic activity — to address a mortality risk of roughly 0.3 percent. Many experts believe that catastrophic risks from advanced A.I. over the next decade are at least this large, suggesting that a comparable mitigation investment could be worthwhile. Existing lives are valued by policymakers at around $10 million each in the United States. To avoid a 1% mortality risk, this value implies a willingness to pay of $100,000 per person — more than 100% of per capita GDP. If the risk is realized over the next two decades, an annual investment of 5% of GDP toward mitigating catastrophic risk could be justified, depending on the effectiveness of such investment. This back-of-the-envelope intuition is supported by the model developed here. In the model, for most of the scenarios and parameter combinations considered, spending at least 1% of GDP annually to mitigate AI risk can be justified even without placing any value on the welfare of future generations.

Charles I. Jones
Graduate School of Business
Stanford University
655 Knight Way
Stanford, CA 94305-4800
and NBER
chad.jones@stanford.edu

Slides are available at https://web.stanford.edu/~chadj/slides-reducerisk.pdf

1. Introduction

The rise of artificial intelligence (A.I.) in recent years is breathtaking, and the pace of progress shows little sign of slowing.[1] Experts including Sam Altman, Dario Amodei, Demis Hassabis, Geoff Hinton, and Ilya Sutskever — among many others — have all highlighted the double-edged nature of A.I.: it can be an incredible technology that raises global living standards, but only if we take care to avoid potentially catastrophic outcomes, either from malicious use or even from a superintelligent A.I. itself. Creating an artificial intelligence that is significantly better than humans at most tasks and that can be scaled up simply by adding more computers could be world changing. A hundred million virtual Einsteins, Von Neumanns, and Doudnas could advance science and technology in a myriad of ways, perhaps leading to nearly free clean energy, an era of economic abundance, and cures for cancer and other health problems (Aschenbrenner, 2024; Amodei, 2024).

However, experts also warn that such a technology could potentially create catastrophic harm: it may be more important than the internet or electricity but also more dangerous than nuclear weapons.[2] One form of risk is from "bad actors": consider a datacenter full of Nobel-caliber A.I. experts in biochemistry that could invent a virus 10 times more deadly and more virulent than smallpox. A more speculative risk could come from the "alien" superintelligence of a future A.I. How would we feel upon learning that an alien spaceship has been detected near Saturn on its way to Earth? Quoting from computer scientist Stuart Russell, coauthor of one of the most popular graduate textbooks on artificial intelligence, "How do we have power over entities more powerful than us, forever?"[3] Grace, Stewart, Sandkühler, Thomas, Weinstein-Raun and Brauner (2024) survey more than 2,500 researchers who have published in "top-tier" A.I. outlets. When asked to quantify the existential risk from A.I., the median estimate was 5% or

[1] For example, in December 2024, OpenAI previewed their o3 model. On PhD-level exams and in computer coding competitions, the model performs significantly better than domain experts, for example by scoring 87.7% on the GPQA PhD-level science benchmark (versus around 70% for in-field PhD students). Just a few months earlier, the o1 model was scoring 78%. See https://www.datacamp.com/blog/o3-openai and Rein et al. (2023).

[2] The case that the existential risk from A.I. merits serious concern has been made by Joy (2000), Bostrom (2002, 2014), Rees (2003), and Yudkowsky et al. (2008). Hendrycks, Mazeika and Woodside (2023) and Ngo, Chan and Mindermann (2023) provide recent overviews.

[3] From a presentation at the CEPR Webinar Series on the Economics of Artificial Intelligence on January 28, 2025.

10% and mean estimates exceeded 10% (depending on the exact question). There is obviously uncertainty about the risk itself, but the fact that experts take it seriously suggests that it is a problem worth considering.

This paper studies a question that at first struck me as too open-ended to be usefully addressed by standard economics: How much should we spend to reduce the existential risks associated with A.I.? But our recent experience with Covid-19 provides a key insight into this question.

Covid-19 was a catastrophic pandemic that killed 1.3 million people in the United States and more than 7 million people worldwide with a mortality rate of around 0.3% (Worldometer, 2024). In response to this catastrophic threat to life, individuals substantially limited economic activity, leading to losses of nearly 4% of GDP in the United States and in many other economies (Goolsbee and Syverson, 2021; Fernández-Villaverde and Jones, 2020).

There are of course many limitations to this analogy, and I do not want to dwell on it. However, it is a quite suggestive benchmark both for the empirical answer it suggests and because it motivates a way to use economic analysis to address the question. We can model the existential risk from A.I. just like we modeled mortality in the Covid-19 pandemic. The mortality risk to people already alive more than justifies spending large amounts to mitigate A.I.'s existential risk. Placing value on the welfare of future generations would of course raise the amount. But the point that we are very likely underinvesting in mitigation is already implied by a selfish perspective.

Section 2 develops a simple model of existential risk and mitigation. Section 3 calibrates the model and derives estimates of optimal mitigation spending as a share of GDP. Across a range of simulations, spending at least 1% of GDP every year over the next decade to mitigate the existential risk associated with A.I. is desirable in most cases, even ignoring future generations. And the average optimal mitigation share across a range of Monte Carlo simulations exceeds a stunning 8% of GDP. Adding a modest value of altruism toward future generations or increasing the potential level of risk substantially increases these numbers. There are scenarios in which one would not want to invest in mitigation. These essentially involve low extinction risks combined with an ineffective technology for reducing risk.

One place where the Covid-19 analogy breaks down is that the pandemic was relatively short and there is little sense that the 4% of GDP we spent mitigating Covid-19 was optimal. The model in this paper is similar to one that Hall, Jones and Klenow (2020) constructed to compute society's willingness to pay to avoid Covid-19 deaths. Assuming a mortality rate of 0.44% from Covid-19 (based on early estimates), that paper found that society would be willing to pay 28% of GDP to avoid the pandemic.

The intuition for large numbers in both that paper and this one is straightforward. Standard estimates of the "value of life" are high. U.S. government agencies implementing safety policies routinely use numbers on the order of $10 million or more for the value of life for an average American today (U.S. Environmental Protection Agency, 2024; U.S. Department of Transportation, 2025). To avoid a mortality risk of 1%, this value implies a willingness to pay of $1\% \times \$10$ million $= \$100,000$. Average GDP per person is around $85,000, so this willingness to pay is more than 100% of GDP. Note that income is an annual number whereas the willingness to pay values the stock of remaining life — perhaps 40 years or more. If the mortality risk is realized once in the next 10 to 20 years, an annual investment of 5–10% of income could be appropriate. This willingness to pay needs to be multiplied by a measure of the effectiveness of the mitigation spending. With that adjustment, the model suggests that spending more than 1% of GDP per year on mitigation is typically justified.

Related literature. Many recent papers emphasize the potentially large benefits of A.I., including Brynjolfsson and McAfee (2014), Aghion, Jones and Jones (2019), Trammell and Korinek (2020), Davidson (2021), Nordhaus (2021), and Erdil and Besiroglu (2023). Brynjolfsson, Korinek and Agrawal (2024) propose a research agenda to study the economics of transformative A.I. Other papers consider potential economic costs in terms of the labor market (Acemoglu and Restrepo, 2022; Autor and Thompson, 2024; Deming et al., 2025).

Jones (2016) considers the tradeoffs between the economic benefits of new technologies and their potential costs in terms of lost lives: as we get richer, life gets more valuable, and it may be optimal to slow or even stop the development of risky new technologies. Jones (2024) studied the application to A.I. explicitly. There, the question was simply: if A.I. development raises growth to 10% per year but comes with a one-time risk of killing everyone, what risks are we willing to take? A surprising finding is that if A.I. reduces mortality, then relatively large risks can be worth taking. The current paper considers a complementary question: assuming A.I. will be developed, how much should we invest to mitigate its risks?

Aschenbrenner (2020) and Aschenbrenner and Trammell (2024) focus on existential risk and 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. Interestingly, they suggest that it may be optimal to accelerate growth to pass through the time of perils more quickly. Growiec and Prettner (2025) provide a thorough review of the work, often outside of economics, related to existential risk and A.I. They develop a rich framework for thinking about the tradeoff between the benefits of A.I. and the risk of existential catastrophe. 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. Much of this work builds on Rosen (1988), Murphy and Topel (2003), Nordhaus (2003), and Hall and Jones (2007) in thinking about how to value lives. Acemoglu and Lensman (2024) show that the optimal adoption of transformative technologies that involve large costs and benefits can be delayed if the costs are irreversible, while Guerreiro, Rebelo and Teles (2023) provide a general analysis of optimal A.I. regulation.

Posner (2004) and Matheny (2007) use value-of-life considerations to argue that the willingness to pay to avoid existential risks of various kinds is high. Ord (2020) and MacAskill (2022) emphasize the potentially trillions of future people whose existence is risked by the decisions of current generations related to nuclear, biological, and A.I. technologies. Shulman and Thornley (forthcoming), like the present paper, observes that this "longtermist" approach is not essential — standard cost-benefit analysis that places no weight on future generations can justify spending large amounts to mitigate existential risk. Nielsen (2024) offers a thoughtful and nuanced discussion of A.I.'s existential risk and how we should approach it.

2. Model

Consider a representative agent deciding how much to consume today versus how much to spend to reduce existential risk. The agent has an exogenous endowment y_t, which can be thought of as growing rapidly because of A.I. The consumption side of the problem is standard: consumption gives utility $u(c_t)$. There is a one-time existential risk realized at the end of the first period — think of periods as T years long, where T is perhaps 5 or 10 years. If the representative agent spends x_t, then the extinction risk is $\delta(x_t)$, a decreasing function of x_t. The agent's (selfish) decision problem is

$$\begin{gathered}\max_{x_t}\ u(c_t) + (1-\delta(x_t))\,\beta\,V_{t+1} \\ s.t.\ \ c_t + x_t = y_t\end{gathered} \tag{1}$$

where $V_{t+1} = \sum_{\tau=0}^{\infty} \beta^{\tau} u(c_{t+1+\tau})$ is future lifetime utility and $c_\tau = y_\tau$ for $\tau > t$.

The first order condition for this problem is

$$u'(c_t) = -\delta'(x_t)\beta V_{t+1} \tag{2}$$

That is, the marginal utility of consumption today equals the marginal benefit of reducing existential risk. This marginal benefit in turn depends on the remaining lifetime utility; we spend for one period to gain a potentially long life.

Multiplying both sides by $x_t/\delta(x_t)$ and rearranging leads to an elegant expression. Let $\eta_{\delta,x} \equiv -\frac{\delta'(x_t)x_t}{\delta(x_t)}$ be the elasticity of extinction risk with respect to spending, and let $s_t \equiv x_t/y_t$ be the fraction of income spent to reduce existential risk. Then the first order condition can be rewritten as

$$\boxed{\frac{s_t}{1-s_t} = \underset{\substack{\text{effectiveness}\\ \text{of spending}\\ >0.01?}}{\eta_{\delta,x}(x_t)} \cdot \underset{\substack{\text{risk to be}\\ \text{mitigated}\\ 1\%?}}{\delta(x_t)} \cdot \underset{\substack{\text{value of}\\ \text{life}\\ >180}}{\beta\frac{V_{t+1}}{u'(c_t)\,c_t}}} \tag{3}$$

The left hand side of this equation is the fraction of GDP devoted to reducing existential risk relative to the fraction devoted to consumption. The equation shows that this is the product of three factors. The first is the elasticity of extinction risk with respect to spending, $\eta_{\delta,x}$. We will impose functional forms and calibrate everything more carefully below, but it is helpful to have some back-of-the-envelope numbers. In this case, it seems plausible that if N people each spend one percent more on reducing existential risk for a T-year period, the probability of extinction falls by at least 0.01%.

The second factor is the amount of extinction risk to be mitigated, $\delta(x_t)$. As discussed in the introduction, estimates of this risk are uncertain, but a reasonable benchmark is a 1% risk over the next two decades.

Finally, the last factor is the future value of life from today's perspective — that is, discounted, converted to consumption units by dividing by the marginal utility of consumption today, and expressed as a ratio to today's consumption: $\tilde{V}_{t+1} \equiv \beta \frac{V_{t+1}}{u'(c_t)\, c_t}$. A standard value of life used by policymakers is \$10 million. Per capita consumption in the United States in 2023 was \$56,000, suggesting a ratio of around 180: the value of an average American's remaining life is around 180 times annual consumption per person. To the extent that A.I. stimulates economic growth, leading to rapid progress in the future, the relevant $\tilde{V}_{t+1}$ is even larger.

The formula in equation (6) is reminiscent of the calculation in Hall, Jones and Klenow (2020). That paper considered society's willingness to pay to avoid the deaths associated with Covid-19 and found that the answer was approximately $\delta \tilde{V}_{t+1}$. This product equals the value of the deaths averted as a share of per capita consumption. There, the mitigation technology was essentially shutting down the economy. Here, we have to consider $\eta_{\delta,x}$ in addition.

Multiplying these lower bounds together suggests that

$$\frac{s}{1-s} \geq 0.01 \times 1\% \times 180 = 1.8\%.$$

That is, it could be optimal to spend at least 1.8% of GDP on reducing existential risk. This back-of-the-envelope calculation ignores the endogeneity of $\eta_{\delta,x}$, $\delta(x_t)$ and $c_t = y_t - x_t$, but as we will see below, the calculation is informative.

Future generations. One way in which the calculation so far might understate optimal mitigation investment is that it ignores the welfare lost by future generations if the existential risk is realized. Incorporating future generations is relatively straightforward in that we can simply augment V_{t+1} with an additional term that captures the extent to which the aggregate welfare of future generations is valued, W_F. That is, the V_{t+1} in the first order condition is replaced by $V_{t+1} + W_F$. While this is straightforward at a mathematical level, it obviously introduces complicated philosophical issues when it comes to calibrating W_F. We discuss this below.

Other existential risks. This paper is framed around the existential risk associated with A.I., but it is related to other existential risks in two ways. First, the general framework could readily be applied to mitigating other kinds of existential risk such as from asteroids or climate change. Second, including other risks could affect our willingness to invest to mitigate A.I.'s risk, as in the competing risks framework of Dow, Philipson and Sala-i-Martin (1999). Such risks could readily be included via the discount factor β. To the extent that these other risks are are small — e.g., see Ord (2020) — the calibration will not be much affected.

3. Functional Forms, Intuition, and Numerical Results

Functional forms. To make additional progress, we assume

$$\delta(x) = (1-\phi)\delta_0 + \phi\delta_0 e^{-\alpha N x} \tag{4}$$

$$u(c) = \bar{u} + \frac{c^{1-\theta}}{1-\theta} \tag{5}$$

With zero mitigation effort, existential risk is δ_0. With infinite mitigation efforts, risk falls to $(1-\phi)\delta_0$; that is, ϕ is the fraction of the risk that can be eliminated by spending. The parameter α governs the effectiveness of total mitigation spending, where the total is Nx: N is the number of people each spending x.

Calibration. To solve the model, we need to calibrate the various parameters, summarized in Table 1. We choose the units of population and GDP so that $N = 1$ and $y_0 = 1$. The parameters of the existential risk function $\delta(x)$ are obviously important, and we consider a wide range of values. The parameter δ_0 is the probability of extinction with zero mitigation effort. We start with a relatively conservative range and consider values between 0% to 2%, with a baseline value of 1%; recall this is a one-time risk that applies over the next decade or so. The parameter ϕ is the fraction of extinction that can be mitigated by spending; $(1-\phi)\delta_0$ is the level of risk that cannot be eliminated. We consider values of ϕ ranging from 0 to 1 with a baseline value of 0.5.

The parameter α is the effectiveness of spending at reducing existential risk. We calibrate this parameter by expressing it in a different way: What fraction of the risk that can be mitigated would be eliminated by investing 100% of GDP for a single year?

Table 1: Baseline Parameter Values

	Parameter	Value	Distribution
Extinction risk, no mitigation	δ_0	1%	Uniform (0%, 2%)
Share that can be eliminated	ϕ	0.5	Uniform (0, 1)
Effectiveness of spending	ξ	0.5	Uniform (0, 0.99)
Value of life	$V_{t+1}/u'(y_t)$	180	Uniform (0.5*180, 1.5*180)
Time of perils (period length)	T	10 years	Uniform (5, 20)
CRRA	θ	2	...
Discount factor	β	0.99^T	...
Value of future generations	W_F	0	...

Let ξ denote this fraction. In the model, the existential risk is realized after one period, but this risk could occur over 5 or 10 or even 20 years. Let T denote the number of years in one period. Then $\alpha N = -T \log(1-\xi) \approx \xi T$.[4] We consider values of ξ from 0 to 0.99, with a baseline value of 0.5.

The remaining parameters in equation (6) are more conventional. The marginal utility of consumption is $u'(c) = c^{-\theta}$; we assume $\theta = 2$. For the baseline value of life with zero mitigation spending, we choose $10 million. As discussed in the introduction, this is a conventional value even ignoring any benefits from A.I. and so is conservative. Given that we choose units so that per capita consumption today is 1, this leads to $V_{t+1}/u'(y_t) = 180$. That is, the remaining life of a 40-year American is worth around 180 years of average consumption.[5]

In the model, the existential risk is realized after one period. How long is that time? Given the advances in A.I. in recent years, we consider a baseline of 10 years and consider robustness to a range from 5 to 20 years.[6] We choose an annual discount factor of

[4] The amount of risk that can be mitigated is $\phi\delta_0$. One year's worth of GDP is such that $x = 1/T$ since a period is T years long. Therefore α solves the equation

$$\delta(1/T) - \delta(\infty) = (1-\xi)\phi\delta_0$$

which gives the solution in the text.

[5] This uses consumption per person in 2023 of around $56,000.

[6] In solving the model, we scale down the value of life by the period length to account for the fact that the risk is realized in T years, by which time life expectancy will be lower.

0.99; with the period length of 10 years $\beta = 0.99^{10}$. Finally, our baseline calculation is entirely selfish and does not consider the value of future generations ($W_F = 0$).

3.1 Analytic Solutions and Intuition

Using the functional form for $\delta(x)$ in (4), optimal spending satisfies

$$e^{\alpha N x_t} = \underset{\substack{\text{effectiveness}\\ \text{term}}}{\alpha N \phi \delta_0} \cdot \underset{\substack{\text{value of life}\\ \text{(in dollars)}}}{\beta \frac{V_{t+1}}{u'(c_t)}} \tag{6}$$

Notice that $u'(c_t) = (y_t - x_t)^{-\theta}$, so the right-hand side of the equation is decreasing in x. Since the left-hand side is increasing in x, there is typically a unique solution to this equation.

For additional intuition, consider the following approximations: $e^{\alpha N x} = 1 + \alpha N x$ which is valid for $\alpha N x$ small, $u'(c) = u'(y - x) = u'(y)$ which is valid for x small. In addition, recall that $\alpha N = -T \log(1 - \xi) \approx \xi T$. With these approximations, the solution in equation (6) can be written as

$$s \equiv \frac{x_t}{y_t} \approx \underset{\substack{\text{WTP = willing-}\\ \text{ness to pay}}}{\phi \delta_0 \beta \frac{V_{t+1}}{u'(y_t) y_t}} - \underset{\substack{\text{effectiveness}\\ \text{of mitigation}}}{\frac{1}{\xi T y_t}} \tag{7}$$

The optimal mitigation share is the sum of two terms. The first term $\phi \delta_0 \beta V_{t+1}/u'(y_t) y_t$ is the "willingness to pay" (WTP): the expected value of the lives that could be lost to the mitigable portion of existential risk, measured as a ratio to per capita GDP (or consumption). For example, start with the value of life as \$10 million for a 40-year old with 40 years to live. Since a period is 10 years, the person only has 30 years to live one period from now, so we use \$7.5m. Including discounting, this value is roughly 120 times a year's per capita consumption (the same as GDP in this model). The baseline existential risk is δ_0 (e.g. 1%) and a fraction ϕ (e.g. 1/2) of that can be mitigated. The willingness to pay if all the risk could be mitigated is 1.2 year's worth of GDP (1% of 120), but since only half that risk can be mitigated, the WTP term equals 0.6, i.e. 60% of a year's GDP. That's the baseline value of the first term.

The second term, $\frac{-1}{\xi T y_t}$, adjusts this WTP for the overall effectivness of spending: ξ is the fraction of the mitigable risk that would be eliminated by spending one year of GDP, we spend for one period which equals T years, and y_t converts this to a share of GDP. In our baseline calibration, $\xi = 1/2$, $T = 10$, and $y_t = 1$. With these values, the second term subtracts off 0.2 or 20%. The net of the first and second terms, then, is $0.60 - 0.20 = 0.40$, suggesting that $s = 40\%$ of GDP! This is so large that the approximations are not valid, but it gives an indication of how large amounts of spending can be optimal.

For a more conservative calibration, suppose the baseline risk δ_0 is half as large at 0.5%. In that case, the WTP term is half as large at 30% of GDP. Subtracting the 0.20 mitigation effectiveness term means that optimal mitigation spending is around 10% of GDP (for T years). As we will see in the next section, the numerical solution for these parameter values is 8.3% of GDP, so the approximation is relatively good.

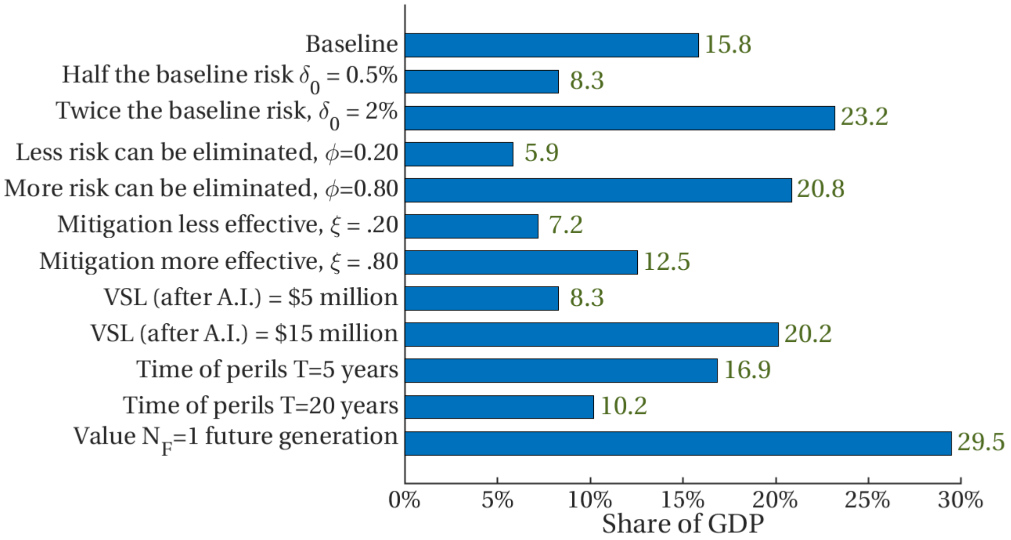
3.2 Numerical Solutions for Specific Parameter Values

Figure 1 shows the optimal share of GDP devoted to reducing existential risk for a range of parameter values, solving equation (6) numerically. At the baseline parameter values, the optimal share is 15.8%. Even if the extinction risk is lower at $\delta_0 = 0.5\%$ — roughly at the Covid-19 level — the optimal share remains high at 8.3%. The intuition for these large spending shares is precisely that given above and revealed by our experience with Covid-19: the value of life for the average American is high — around $10 million and thus more than 120x our annual consumption per person. It is therefore worth spending a large fraction of GDP over the next decade to reduce existential risk.

The remaining bars in Figure 1 vary one parameter at a time away from the baseline values. In all cases, optimal mitigation remains high — above 5% of GDP in all cases. For example, the case of $\phi = 0.20$ supposes that only 20% of the extinction risk can be eliminated, even with infinite spending (as opposed to 50% in the baseline). The case of $\xi = 0.20$ supposes that investing a year's worth of GDP only reduces the mitigable risk by 20% (again as opposed to 50% in the baseline). If the value of life is only $5 million instead of $10 million, optimal mitigation is 8.3% of GDP.

The cases of $T = 5$ and $T = 20$ show how the results vary with the length of our "time of peril." If we only have 5 years until the existential risk is realized, the model says we should spend an even larger fraction of GDP on mitigation (16.9%). If the risk is further

Figure 1: Optimal Spending to Reduce Existential Risk



away at 20 years, the optimal share per year falls to 10.2%: more years of investment can make up for less investment per year.

Finally, the last case in Figure 1 introduces a degree of altruism toward future generations that is missing from our selfish baseline. To understand the calibration in this case, we suppose that $W_F = N_F V_F$, where V_F is the average value of life for future generations and N_F is the effective number of future people being considered (relative to the current population of $N = 1$). If the existential risk associated with A.I. is not realized, then future generations could be much better off, both because their consumption is high and, importantly, because their life expectancy may increase (Jones, 2024). Nevertheless, to be conservative, we set $V_F = 2V_{t+1}$. Recall that V_{t+1} is the value of remaining life for an average American, i.e. someone about 40 years old. Since V_F is the value of life for a newborn, we multiply by two. And for the number of future generations that benefit from A.I., we consider $N_F = 1$. That is, the effective number of future people we are considering is equal in size to the current population. With these parameter values, the results will be equivalent to tripling the value of V_{t+1}. With this quite limited degree of altruism toward future generations, optimal mitigation spending rises from 15.8% to 29.5%.

The bottom line is that in all of these variations, it is optimal to invest more than 5%

Table 2: Optimal Mitigation Results

	Mitigation spending share (percent)	Remaining x-risk, $\delta(x)$ (percent)	Mitigation elasticity $\eta(x)$	Fraction mitigated $\frac{\delta(0)-\delta(x)}{\delta(0)-\delta(\infty)}$
Baseline	15.8	0.67	0.275	0.666
Half the baseline risk, $\delta_0 = 0.5\%$	8.3	0.39	0.207	0.438
Twice the baseline risk, $\delta_0 = 2\%$	23.2	1.20	0.268	0.800
Less risk can be eliminated, $\phi = 0.20$	5.9	0.93	0.058	0.333
More risk can be eliminated, $\phi = 0.80$	20.8	0.39	0.701	0.764
Mitigation less effective, $\xi = 0.20$	7.2	0.93	0.074	0.148
Mitigation more effective, $\xi = 0.80$	12.5	0.57	0.237	0.867
VSL after A.I. = $5 million	8.3	0.78	0.207	0.438
VSL after A.I. = $15 million	20.2	0.62	0.277	0.753
Time of perils T=5 years	16.9	0.78	0.209	0.443
Time of perils T=20 years	10.2	0.62	0.277	0.757
Value $N_F = 1$ future generation	29.5	0.56	0.234	0.870

of GDP each year to mitigate existential risk.

Table 2 reports additional results for these cases. For example, in the baseline case, two-thirds of the mitigable risk is mitigated by spending so that the remaining existential risk equals 0.67%. The third column in the table shows that the mitigation elasticity $\eta(x) \equiv -\frac{\delta'(x)x}{\delta(x)}$ is typically around 0.2 but can fall as low as 0.058 in these scenarios; this elasticity can be compared to the intuition given earlier in equation (3).

When would we not invest? One way to study the robustness of these results is to consider the scenarios that would lead to low or even zero investment toward mitigating existential risk. To see this, return to equation (2). In particular, it is optimal to not invest in mitigation when

$$u'(y_0) > -\delta'(0)\beta V_{t+1}$$

that is, when the marginal utility of consumption today, even when doing no mitigation so that $c_t = y_0$, exceeds the marginal benefit of mitigation. Plugging in our functional form for $\delta(x)$ as in equation (6) gives

$$1 \; > \; \alpha N \; \cdot \; \phi\delta_0\beta\frac{V_{t+1}}{u'(y_0)} \; \approx \; \underset{\substack{\text{effectiveness}\\ \text{of spending}}}{\xi\, T} \quad \cdot \quad \underset{\substack{\text{WTP}\\ \text{= EV of lives}\\ \text{lost to x-risk}}}{\phi\delta_0\beta\tfrac{V_{t+1}}{u'(y_0)}} \tag{8}$$

where the approximation uses the fact that $\alpha N = -T\log(1-\xi) \approx \xi T$.

The last part of equation (8) is the product of the same two terms we saw earlier in the approximation in equation (7). The first, ξT, is the overall effectiveness of spending. The second, $\phi\delta_0\beta V_{t+1}/u'(y_0)$, is the "willingness to pay" (WTP) — the expected value of the lives that could be lost to the mitigable portion of existential risk, measured as a ratio to per capita GDP (since y_0=1).

Simplifying the expression, optimal mitigation investment would be zero only if

$$\xi\, T \; \cdot \; \text{WTP} < \; 1$$

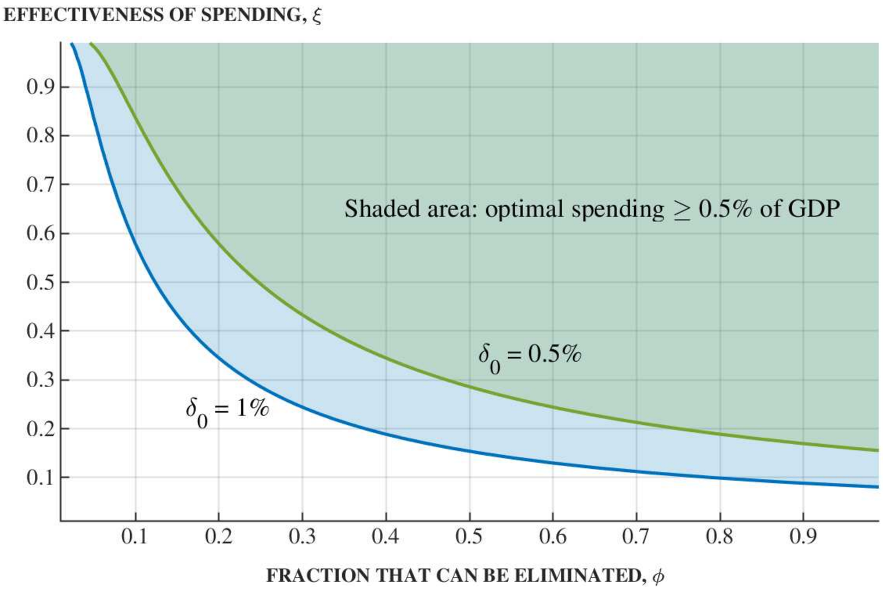
For our baseline parameter values, $\xi = 1/2$, $T = 10$, and WTP$= 60\%$ of GDP, so the left side takes a value of 3, far exceeding one, which is why optimal investment is large.

What could make the left side small? Clearly ξ and ϕ could be small if mitigation is not effective. For example, suppose the fraction of risk that can be mitigated (ϕ) is 5x lower — which lowers WTP by 5x. Or suppose mitigation spending is 5x less effective, reducing ξ by a factor of 5. Both of these would reduce the LHS to 3/5, below one, so that zero investment would be optimal. Alternatively, if the baseline risk δ_0 were very small — say only 0.1% instead of 1%, then the left side would be 10x smaller and optimal investment would again be zero. These are the kind of scenarios that deliver zero mitigation investment — we will see examples of this in the Monte Carlo exercises shortly.

When is optimal spending $\geq$ 0.5% of GDP? Motivated by this intuition, Figure 2 shows the combinations of key effectiveness parameters that lead to optimal spending of at least 0.5% of GDP.[7] The scenarios that lead to low mitigation investment essentially

[7] One can draw a similar plot where the cutoff is zero instead; that plot looks extremely similar to Figure 2: the gradient of spending with respect to the key parameter values is very steep, consistent with the results already shown in Figure 1.

Figure 2: When is optimal spending $\geq$ 0.5% of GDP?



Note: The figure displays combinations of parameter values that lead to optimal spending of at least 0.5% of GDP. The parameter ξ answers the question "What fraction of the risk that can be mitigated would be eliminated by investing 100% of GDP for one year?"

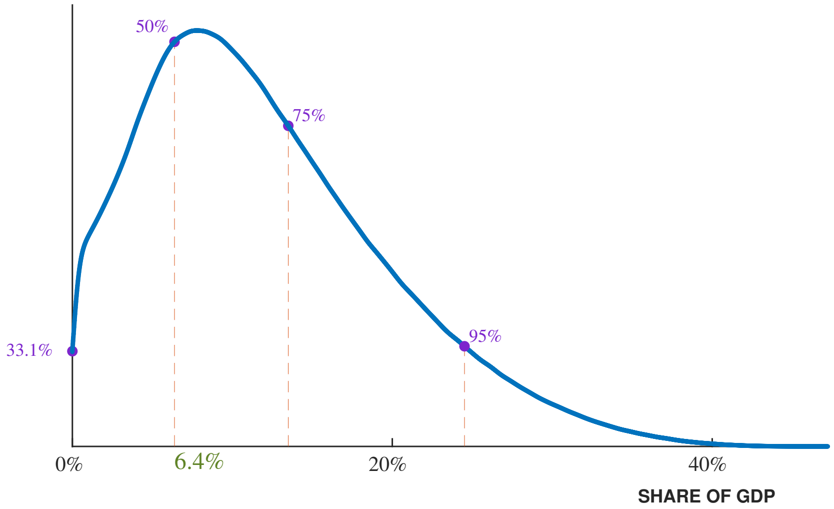
involve low extinction risks combined with an ineffective technology for reducing risk.

3.3 Monte Carlo Results

There is obviously a great deal of uncertainty regarding the appropriate values of the various parameters. To incorporate this uncertainty — and to illustrate the robustness of our main result — we consider a large number of simulations with different parameter values. The distributions we choose for the key parameters are summarized in the last column of Table 1. For example, we assume the extinction risk is uniformly distributed between 0% and 2%, and the fraction of risk that can be mitigated is uniformly distributed between 0% and 100%. For the value of life, we take a uniform distribution with a minimum of $5 million (e.g. if one believes existing estimates overstate the value of life) and a maximum of $15 million (e.g. incorporating some of the benefits of A.I., such as longer lives) so that the mean matches our baseline value of $10 million.

Figure 3 shows the distribution of optimal mitigation across 10 million Monte Carlo

Figure 3: Monte Carlo Results for Optimal Spending



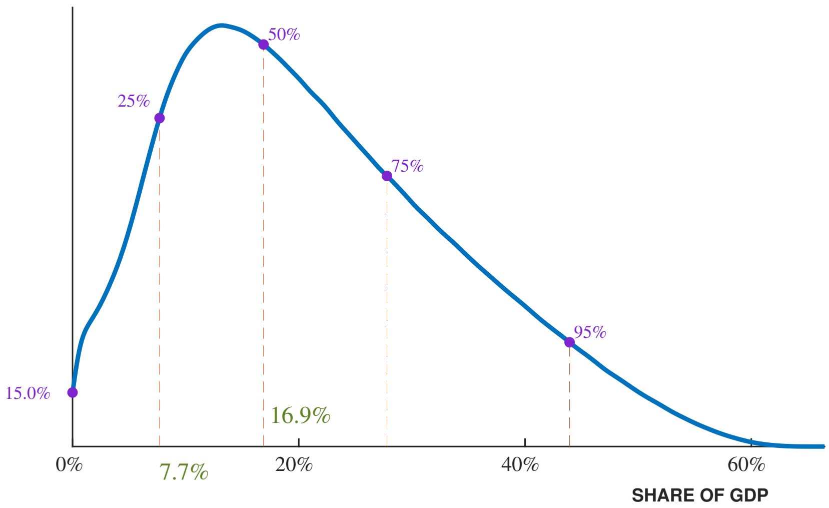
Note: The figures plot the distribution of optimal mitigation spending as a share of GDP across 10 million simulations using the distributions of parameters in Table 1. For example, δ_0 is `Uniform[0,2%]`.

simulations for the baseline case of a selfish allocation that places no value on future generations. Summary statistics for these distributions are given in Table 3 — that is, for the distributions reported in Table 1. Figure 4 shows the distributions for two other cases: a scenario that allows for modest altruism toward a future that is the size of the present ($N_F = 1$) and a scenario that allows for a higher maximum extinction risk (δ_0 is `Uniform[0,10%]`).

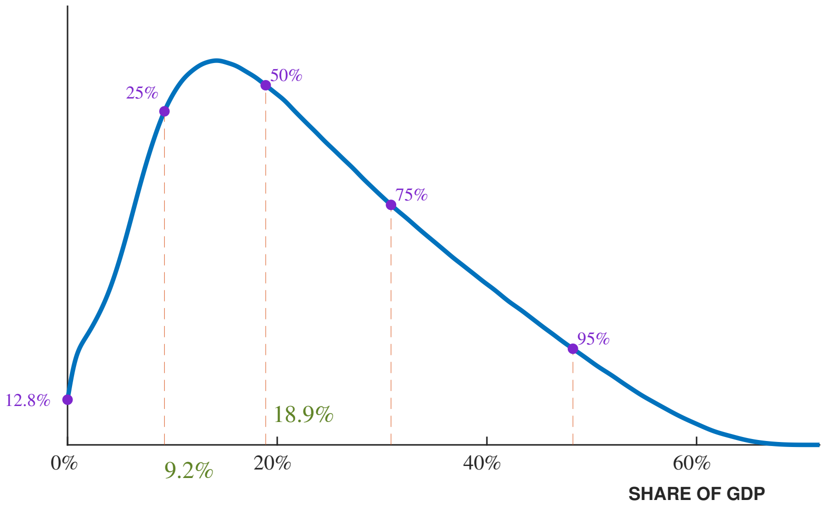
In all three of the Monte Carlo graphs, a significant share of the runs — between 12% and 33% — lead to zero mitigation investment. Recall that our uniform distributions allow for the possibility that the existential risk is very small or that it is very hard to mitigate. As discussed in the previous section, those are the scenarios that lead to zero investment.

Nevertheless, the main conclusion to be drawn from the simulation results is that it is optimal to spend a remarkably large share of GDP to mitigate existential risk. Even without considering the value of future generations, the average optimal mitigation share is 8.1% of GDP. When we consider some modest altruism toward future generations or a great maximum extinction risk, the optimal spending shares are even higher.

Figure 4: Monte Carlo Results: Robustness



(a) Modest altruism ($N_F = 1$)



(b) Higher possible risk (δ_0 is `Uniform[0,10%]`)

Note: See notes to Figure 3.

Table 3: Summary Statistics for Monte Carlo Simulations

	Selfish baseline $(N_F = 0)$ $\delta_0 \sim \texttt{Uniform[0,2\%]}$	Modest altruism $(N_F = 1)$	Higher risk $(N_F = 0)$ $\delta_0 \sim \texttt{Uniform[0,10\%]}$
Optimal share, mean	8.1%	18.4%	20.7%
Fraction with $s_t = 0$	33.1%	15.0%	12.8%
Fraction with $s_t \geq 1\%$	65.1%	84.2%	86.5%

Note: The table shows summary statistics for the distributions of optimal mitigation spending across 10 million simulations based on the distributions of parameters in Table 1, except as noted.

4. Conclusion

Just as it made sense to spend around 4% of GDP to limit deaths from the Covid-19 pandemic (with a 0.3% mortality risk), it may be worthwhile to spend large amounts to mitigate a potential catastrophic risk from artificial intelligence. Standard estimates of the value of life in the United States are around $10 million. This implies a willingness to pay of $100,000 per person to avoid a 1% mortality risk, more than 100% of GDP. If the risk is realized over the next decade, the willingness to pay exceeds 10% of annual GDP. This willingness to pay must be multiplied by the effectiveness of mitigation spending. For example, even if only 50% of the risk can be mitigated and even if mitigation spending is not that effective (e.g. spending a full year of GDP would reduce this risk by only 50%), then optimal mitigation spending in the model is, stunningly, more than 8% of GDP each year. Note that this calculation is selfish in the sense that it puts zero value on future generations. Even a selfish perspective suggests that we might spend significant amounts to mitigate existential risk.

More broadly, there are many related questions that merit further study. For example, there are clearly huge externalities to the actions taken by the A.I. labs, and the equilibrium allocation without intervention is almost certainly Pareto inefficient. What policies are called for? Should we tax GPUs and use the revenue to fund safety research?

References

Acemoglu, Daron and Pascual Restrepo, "Tasks, Automation, and the Rise in U.S. Wage Inequality," *Econometrica*, 2022, *90* (5), 1973–2016.

__ and Todd Lensman, "Regulating Transformative Technologies," *American Economic Review: Insights*, September 2024, *6* (3), 359–76.

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.

Amodei, Dario, "Machines of Loving Grace: How AI Could Transform the World for the Better," October 2024.

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

__ , "Situational Awareness: The Decade Ahead," June 2024.

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

Autor, David and Neil Thompson, "Does automation replace experts or augment expertise? The answer is yes," August 2024. Schumpeter Lecture, European Economic Association.

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.

__ , Anton Korinek, and Ajay Agrawal, "The Economics of Transformative AI: A Research Agenda," December 2024.

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

Deming, David J, Christopher Ong, and Lawrence H Summers, "Technological Disruption in the Labor Market," Working Paper 33323, National Bureau of Economic Research January 2025.

Dow, William H., Tomas J. Philipson, and Xavier Sala-i-Martin, "Longevity Complementarities under Competing Risks," *American Economic Review*, December 1999, *89* (5), 1358–1371.

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

Fernández-Villaverde, Jesús and Charles I. Jones, "Macroeconomic Outcomes and COVID-19: A Progress Report," *Brookings Papers on Economic Activity*, Fall 2020, pp. 111–166.

Goolsbee, Austan and Chad Syverson, "Fear, lockdown, and diversion: Comparing drivers of pandemic economic decline 2020," *Journal of Public Economics*, 2021, *193*, 104311.

Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner, "Thousands of AI Authors on the Future of AI," 2024.

Growiec, Jakub and Klaus Prettner, "The Economics of p(doom): Scenarios of Existential Risk and Economic Growth in the Age of Transformative AI," January 2025. unpublished manuscript.

Guerreiro, Joao, Sergio Rebelo, and Pedro Teles, "Regulating Artificial Intelligence," Working Paper 31921, National Bureau of Economic Research November 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).

Hendrycks, Dan, Mantas Mazeika, and Thomas Woodside, "An Overview of Catastrophic AI Risks," 2023.

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

__ , "The AI Dilemma: Growth versus Existential Risk," *American Economic Review: Insights*, December 2024, *6* (4), 575–90.

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.

Matheny, Jason G., “Reducing the Risk of Human Extinction,” *Risk Analysis*, 2007, *27* (5), 1335–1344.

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.

Nielsen, Michael, “How to be a wise optimist about science and technology?,” December 2024. https://michaelnotebook.com/optimism/index.html.

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.

Rein et al., David, “GPQA: A Graduate-Level Google-Proof Q&A Benchmark,” 2023.

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

Shulman, Carl and Elliott Thornley, “How Much Should Governments Pay to Prevent Catastrophes? Longtermism’s Limited Role,” in Jacob Barrett, Hilary Greaves, and David Thorstad, eds., *Essays on Longtermism*, Oxford University Press, forthcoming.

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

U.S. Department of Transportation, “Departmental Guidance on Valuation of a Statistical Life in Economic Analysis,” Technical Report 2025. Accessed January 10, 2025.

U.S. Environmental Protection Agency, "Mortality Risk Valuation," Technical Report 2024. Accessed December 30, 2024.

Worldometer, "Coronavirus Tracker," 2024. accessed December 29, 2024.

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