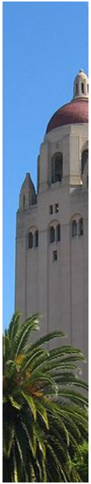


A.I. and Existential Risk

Chad Jones
Stanford GSB

March 2025

Amazing progress in A.I.

- OpenAI's Deep Research, o1pro, Anthropic, Deepmind
  - Coding, math, browsing the internet to write reports
  - Protein folding, understanding DNA, new materials
- Scaling compute + algorithms $= \sim$10x each year
- Huge, exciting opportunities
- But also potentially large risks...
  - Highlighted by many experts (Hinton, Hassabis, Altman, Amodei, etc.)

Can we use economic analysis to think about the serious risks?

Two Versions of Existential Risk

- Bad actors:
    - Could use Claude/GPT-6 to cause harm
    - E.g. design a Covid virus that is 10x more lethal and takes 3 weeks for symptoms
    - Nuclear weapons mangeable because so rare; if every person had them...
- Alien intelligence:
    - How would we react to a spaceship near Jupiter on the way to Earth?
    - “How do we have power over entities more powerful than us, forever?” (Stuart Russell)

Outline

- Quick review of “The A.I. Dilemma” (2024 *AERI*)

- How much should we spend to reduce existential risk?
  - Covid-19 example
  - Using VSL (value of a statistical life) numbers
  - Model and calibration
  - Monte Carlo simulations to incorporate uncertainty regarding risk and effectiveness of mitigation

Even a selfish perspective suggests we are underinvesting in A.I. safety

Related Literature

- A.I. and Growth
  - Brynjolfsson and McAfee (2014), Aghion, Jones, and Jones (2019), Korinek and Trammell (2020), Nordhaus (2021), Growiec and Prettner (2025)
  - Brynjolfsson, Korinek, and Agrawal (2024)
- Costs of A.I.?
  - Acemoglu and Restrepo (2022), Autor, Thompson, and Ong (2024)
  - Jones (2016), Aschenbrenner (2024), Aschenbrenner and Trammell (2024)
- Catastrophic risks
  - Posner (2004), Matheny (2007), Ord (2020), MacAskill (2022), Shulman and Thornley (2025), Nielsen (2024)

A Thought Experiment (Jones, 2024 AERI)

- AGI more important than electricity, but more dangerous than nuclear weapons?
- The Oppenheimer Question:
  - If nothing goes wrong, AGI accelerates growth to 10% per year
  - But a one-time small chance that A.I. kills everyone
  - Develop or not? What risk are you willing to take: 1%? 10%?

What does standard economic analysis imply?

Findings:

- Log utility: Willing to take a 33% risk!

  (Maybe entrepreneurs are not very risk averse?)

- More risk averse ($\gamma = 2$ or 3), risk cutoff plummets to 2% or less
  - Diminishing returns to consumption
  - We do not need a 4th flat screen TV or a 3rd iphone.
    Need more years of life to enjoy already high living standards.

- But 10% growth $\Rightarrow$ cure cancer, heart disease
  - Even $\gamma = 3$ willing to take large risks (25%) to cut mortality rates in half
  - Each person dies from cancer or dies from A.I. Just total risk that matters...
  - True even if the social discount rate falls to zero

How much should we spend to reduce A.I.'s catastrophic risk? (Jones 2025)

How much should we spend to reduce A.I.'s catastrophic risk? (Jones 2025)

- Covid pandemic: "spent" 4% of GDP to mitigate a mortality risk of 0.3%
  - A.I. risk is at least this large $\Rightarrow$ spend at least this much?
  - Are we massively underinvesting in mitigating this risk?

How much should we spend to reduce A.I.'s catastrophic risk? (Jones 2025)

- Covid pandemic: "spent" 4% of GDP to mitigate a mortality risk of 0.3%
  - A.I. risk is at least this large $\Rightarrow$ spend at least this much?
  - Are we massively underinvesting in mitigating this risk?

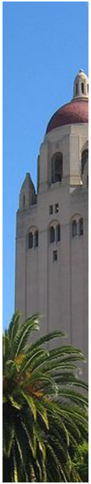
- Better intuition
  - VSL = $10 million
  - To avoid a mortality risk of 1% $\Rightarrow$ WTP = $1\% \times \$10$ million $= \$100,000$
  - This is more than 100% of a year's per capita GDP
  - Xrisk over two decades $\Rightarrow$ **annual investment of 5% of GDP**

- Large investments worthwhile, even with no value on future generations

How much should we spend to reduce A.I.'s catastrophic risk? (Jones 2025)

- Covid pandemic: "spent" 4% of GDP to mitigate a mortality risk of 0.3%
  - A.I. risk is at least this large $\Rightarrow$ spend at least this much?
  - Are we massively underinvesting in mitigating this risk?
- Better intuition
  - VSL = $10 million
  - To avoid a mortality risk of 1% $\Rightarrow$ WTP = $1\% \times \$10$ million $= \$100,000$
  - This is more than 100% of a year's per capita GDP
  - Xrisk over two decades $\Rightarrow$ **annual investment of 5% of GDP**
- Large investments worthwhile, even with no value on future generations

Incomplete so far: how effective is mitigation?



Model

Model

- Setup
  - One-time existential risk at probability $\delta(x)$
  - One-time investment x_t to mitigate the risk ($\delta'(x) < 0$)
  - Exogenous endowment y_t (grows rapidly via A.I.)
- Optimal mitigation:

$$\max_{x_t} u(c_t) + (1 - \delta(x_t))\, \beta\, V_{t+1}$$

$$s.t. \;\; c_t + x_t = y_t$$

$$V_{t+1} = \sum_{\tau=0}^{\infty} \beta^{\tau} u(y_{t+1+\tau}) \qquad \text{(consume } y_t \text{ in future)}$$

Optimal Mitigation

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

- Let $\eta_{\delta,x} \equiv -\frac{\delta'(x_t)x_t}{\delta(x_t)}$ and $s_t \equiv x_t/y_t$

$\frac{s_t}{1-s_t}$	$=$	$\eta_{\delta,x}$	$\cdot$	$\delta(x_t)$	$\cdot$	$\beta \frac{V_{t+1}}{u'(c_t)\, c_t}$
		effectiveness of spending $> 0.1?$		risk to be mitigated $0.1\%?$		value of life > 180

- Taking the smallest numbers:
$$\frac{s}{1-s} \geq 0.1 \times 0.1\% \times 180 = 1.8\%.$$

Additional considerations

- Future generations
  - So far, we place no value on future generations — selfish perspective
  - Easily included: add welfare of future generations W_F to V_{t+1}
- Other existential risks
  - Framework applied to A.I. but can be used to study other risks
  - Competing risks: nuclear war, asteroid impact — include in β

Functional forms

- Existential risk: $$\delta(x) = (1-\phi)\delta_0 + \phi\delta_0 e^{-\alpha N x}$$
  - δ_0 is the risk without mitigation
  - ϕ is the share of the risk that can be eliminated by spending
  - α is the effectiveness of spending
  - N is the number of people each spending x
  - With infinite spending, risk falls to $(1-\phi)\delta_0$
- To calibrate α:

$$\alpha N = -T \log(1-\xi) \approx \xi T$$

ξ is the share of the risk that can be eliminated by spending 100% of GDP for one year

T is "time of perils" = years until risk gets realized (period length)

Calibration

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

	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**	purely selfish for now

Analytic Results and Intuition

- Using the functional forms:

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

Notice that $u'(c_t) = (y_t - x_t)^{-\theta}$, so RHS is decreasing in x.

- Using approximations:

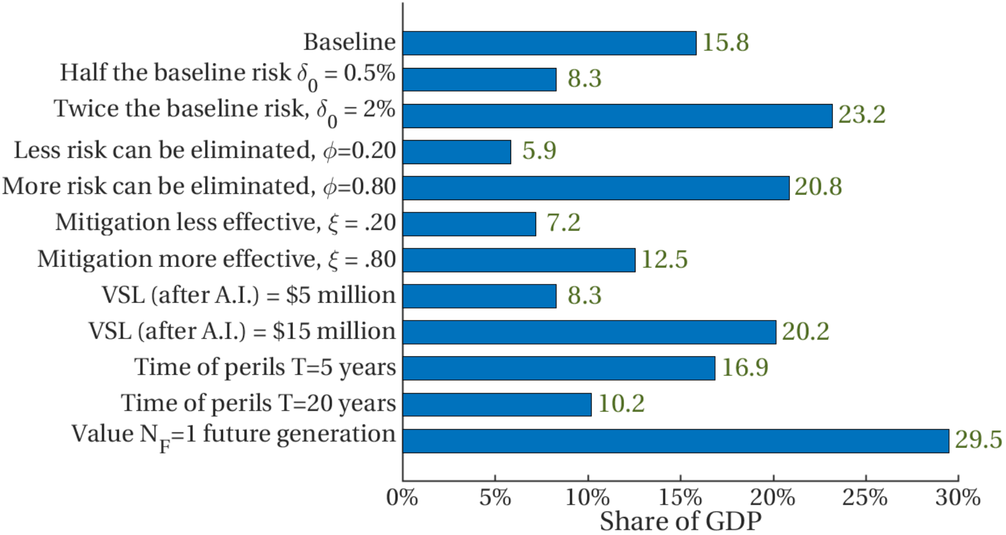
$$s \equiv \frac{x_t}{y_t} \approx \underbrace{\phi \delta_0 \beta \frac{V_{t+1}}{u'(y_t) y_t}}_{\text{WTP = willingness to pay}} - \underbrace{\frac{1}{\xi T y_t}}_{\text{effectiveness of mitigation}}$$

Intuition

$$s \equiv \frac{x_t}{y_t} \approx \underbrace{\phi\delta_0\beta\frac{V_{t+1}}{u'(y_t)y_t}}_{\text{WTP = willingness to pay}} - \underbrace{\frac{1}{\xi T y_t}}_{\text{effectiveness of mitigation}}$$

- WTP term (intuition from an early slides using VSL):
  - $T = 10$, so 40 year old has 30 years remaining$\Rightarrow$ VSL term = 120x consumption
  - $\phi = 1/2$ and $\delta_0 = 1\%$
  - WTP is $0.5 \times 1\% \times 120 = 60\%$ of GDP!
- Mitigation term: $\xi = 1/2$, $T = 10$, and $y_t = 1$ subtracts off 20%
- So approximation is $0.60 - 0.20 = 0.40$, suggesting $s = 40\%$ of GDP!
  - Alternative: $\delta_0 = 0.5\% \Rightarrow s = 10\%$ of GDP, very close to correct 8.3%

Optimal Spending to Reduce Existential Risk



When should we not invest in mitigation?

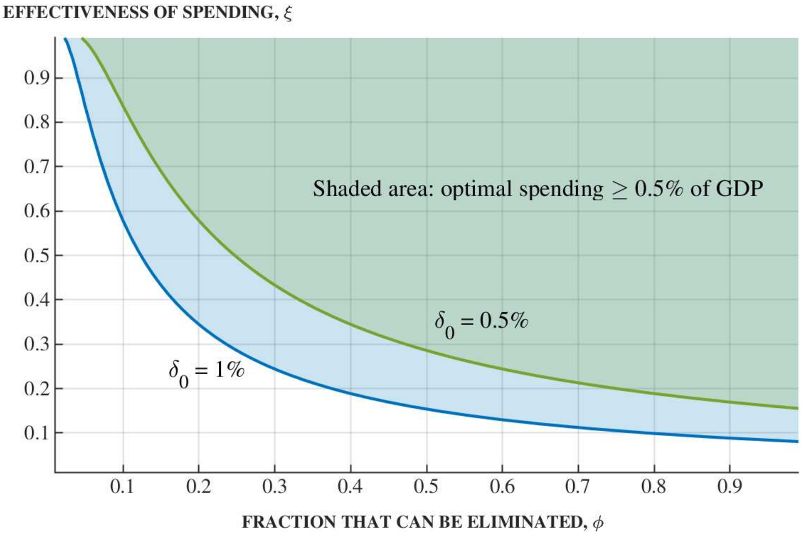
- From FOC: Do not invest if $u'(y_0) > -\delta'(0)\beta V_{t+1}$
- Using functional forms and approximations:

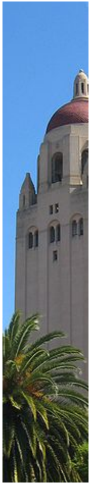
$$1 > \alpha N \cdot \phi\delta_0\beta\frac{V_{t+1}}{u'(y_0)} \approx \underbrace{\xi T}_{\text{effectiveness of spending}} \cdot \underbrace{\phi\delta_0\beta\frac{V_{t+1}}{u'(y_0)}}_{\text{WTP = EV of lives lost to x-risk}}$$

$$\implies \xi T \cdot \text{WTP} < 1$$

- $\xi = 1/2$, $T = 10$, and WTP $= 60\%$ of GDP, LHS = 3
  - But ϕ or ξ or δ_0 $\Rightarrow$ 5x smaller $\Rightarrow$ invest zero (Little risk, or not much can be done)

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

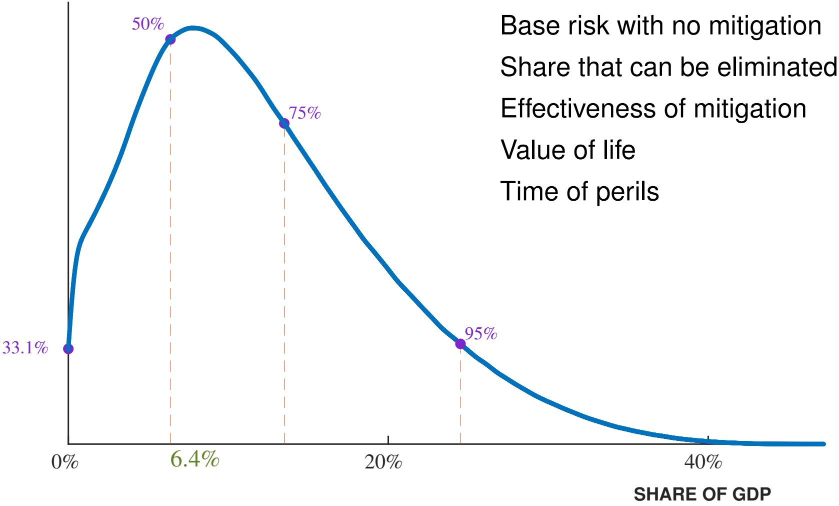




Monte Carlo Results

10 million simulations

Optimal Mitigation: Monte Carlo Simulation

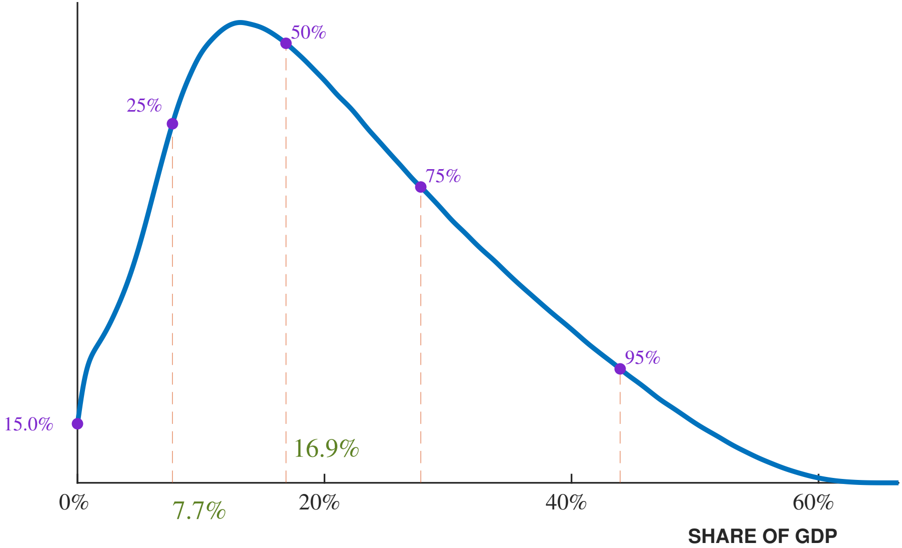


Uniform distributions over:

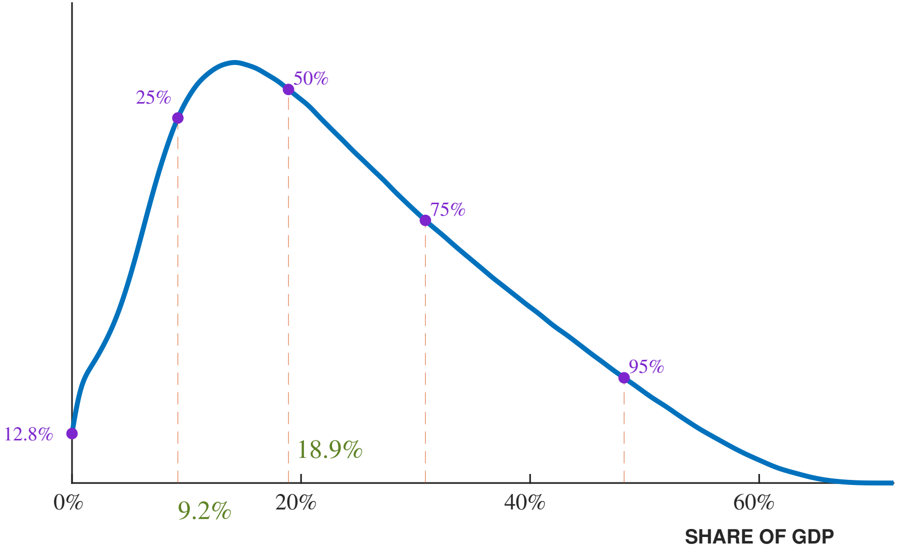
Base risk with no mitigation	0 – 2%
Share that can be eliminated	0 – 100%
Effectiveness of mitigation	0 – 99%
Value of life	\$5m – \$15m
Time of perils	5 – 20 years

Mean = 8%. 65% of runs have $s \geq 1\%$

Modest Altruism toward a Same-Size Future$(N_F = 1)$



Higher Potential Risk (δ_0 is `Uniform[0,10%]`)



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%

Concluding Questions

- How large is the catastrophic risk from A.I.?
  - How much are we currently spending to mitigate A.I. risk?
  - What evidence is there on the effectiveness of mitigation spending?
- How should we think about A.I. competition and race dynamics?
- How can we get A.I. labs to internalize the x-risk externalities?
  - Should we tax GPUs and use the revenue to fund safety research?