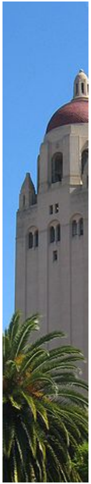


The A.I. Dilemma:
Growth versus Existential Risk

Chad Jones

Stanford GSB

June 2024

The Costs and Benefits of A.I.

- A.I. experts emphasize astounding potential benefits and costs:
  - Benefit: Faster economic growth. Singularity? (it is possible!)
  - Cost: Existential risk — some probability of human extinction
- Not taking a stand on how likely these are
  - Key is that they are highly correlated
- Should we shut down A.I. research or celebrate it?

Outline

- Simple model: Highlight basic considerations
  - Intuitive solution
  - Requires calibrating the existential risk
- Richer model
  - Existential risk cutoff — no need to calibrate the risk itself
  - Singularity?
  - Mortality improvements
  - Longtermism

Cannot provide a firm answer. But models highlight interesting and surprising considerations.

Literature

- Existential risk: Joy (2000), Bostrom (2002, 2014), Rees (2003), Posner (2004), Yudkowsky et al (2008), Ngo et al (2023)
- A.I. and growth: Aghion et al (2019), Trammell and Korineck (2020), Davidson (2021), Nordhaus (2021), Acemoglu and Lensman (2023)
- Life and growth: Jones (2016), Aschenbrenner and Trammell (2024)
- Value of life: Rosen (1988), Murphy and Topel (2003), Nordhaus (2003), Hall and Jones (2007), Martin and Pindyck (2015, 2020)

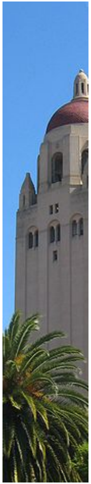
Aghion, Jones, and Jones (2019): A.I. and Growth

- Automation has been ongoing for 200 years, but no growth speedup?
    - Acemoglu-Restrepo task model
    - Baumol cost disease bottlenecks
    - Growth constrained by the tasks at which people are essential

- Romer nonrivalry of ideas means world is characterized by increasing returns
    - Less than complete automation can make share of accumulable factors > 1
    - What if A.I. can produce ideas?

      Virtuous circle: machines $\Rightarrow$ ideas $\Rightarrow$ machines $\Rightarrow$ ideas

    - Growth can speed up



Simple Model

Economic Environment

- Choose T = how intensively to use A.I. (e.g. “how many years”)
  - Consumption: $c = c_0 e^{gT}$ — growth at exogenous rate g, e.g. 10% per year
  - Existential risk: Probability of survival is $S(T) \equiv e^{-\delta T}$.
- Simplify so the model is essentially static:
  - All growth and x-risk occurs immediately
  - If survive, consume constant c_T forever
- N people ⇒ social welfare

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

Optimal Use of the A.I.

- Choose $T \geq 0$ to maximize expected social welfare:

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

- Optimal T^* $\Rightarrow$ use the A.I. as long as

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

$$\Rightarrow \boxed{v(c) \leq \frac{g}{\delta}}$$

- Doesn't depend on N or ρ: All people enjoy both the benefits and the costs forever

Intuition

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

- $v(c) \equiv u(c)/u'(c)c$ = value of a year life life, measured in years of consumption
  - In U.S. today: VSLY≈$250k and $c \approx \$40k \Rightarrow v(c_{us,today}) \approx 6$
  - An average year of life is worth 6 years of per capita consumption

- Call g/δ the A.I. Benefit-Cost (AIBC) ratio

Use the A.I. until the Value of Life $v(c)$ rises to equal the AIBC ratio

CRRA Utility

- Assume

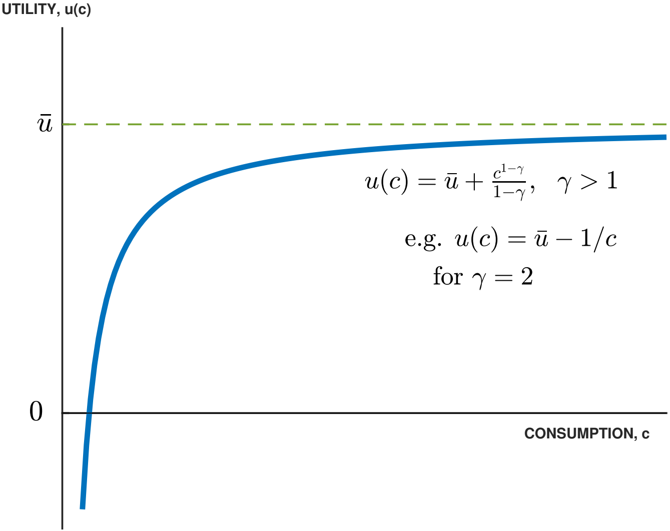
$$u(c) = \begin{cases} \bar{u} + \frac{c^{1-\gamma}}{1-\gamma} & \text{if } \gamma \neq 1 \\ \bar{u} + \log c & \text{if } \gamma = 1 \end{cases}$$

- The value of life is given by

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

– *increases with c for $\gamma \geq 1$*

Bounded flow utility when $\gamma > 1$



$$u(c) = \bar{u} + \frac{c^{1-\gamma}}{1-\gamma}, \quad \gamma > 1$$

e.g. $u(c) = \bar{u} - 1/c$

for $\gamma = 2$

Quantification

- Calibrating key parameters:
  - Growth: $g = 10\%$. High, but taking seriously the most optimistic claims
  - Existential risk: $\delta = 1\%$ or 2%. Useful for illustrating a point
- Recall $v(c_{us,today}) = 6$
  - Normalization: $c_0 = 1$ (choose units)

Consumption and Existential Risk: $\delta = 1\%$

- $g = 10\% \;\Rightarrow\; AIBC = 10 \;\Rightarrow\; v(c^*) = 10$
  - ○ Recall $v(c_{us,today}) = 6$

- **Log utility:** $v(c) = \bar{u} + \log c$

  $\Rightarrow$ $\log c$ rises by 4

Consumption and Existential Risk: $\delta = 1\%$

- $g = 10\% \;\Rightarrow\; AIBC = 10 \;\Rightarrow\; v(c^*) = 10$
  - Recall $v(c_{us,today}) = 6$
- **Log utility:** $v(c) = \bar{u} + \log c$

  $\Rightarrow$ $\log c$ rises by 4
  - $\exp(4) \approx 55$
  - At $g = 10\%$ this takes $T^* = 40$ years
  - $S(T^*) = \exp(-.01 \times 40) \approx 0.67$

Quantitative Results from the Simple Model

γ	c^*	T^*	Exist.Risk
1	54.60	40.0	0.33

With log utility, run the A.I. for 40 years: consumption rises by a factor of 55 — roughly the factor by which U.S. has grown in 2000 years — in exchange for a 1 in 3 chance of extinction!

Consumption and Existential Risk: $\delta = 1\%$

- $g = 10\% \;\Rightarrow\; AIBC = 10 \;\Rightarrow\; v(c^*) = 10$
  - Recall $v(c_{us,today}) = 6$
- **CRRA $\gamma = 2$:** $v(c) = \bar{u} \cdot c - 1$
  - c rises by 100x less: 57% vs. 55x
  - Run the A.I. for $T^* = 4.5$ years
  - $S(T^*) = \exp(-.01 \times 4.5) \approx 0.96$

Quantitative Results from the Simple Model

γ	c^*	T^*	Exist.Risk
1	54.60	40.0	0.33
2	1.57	4.5	0.04
3	1.27	2.4	0.02

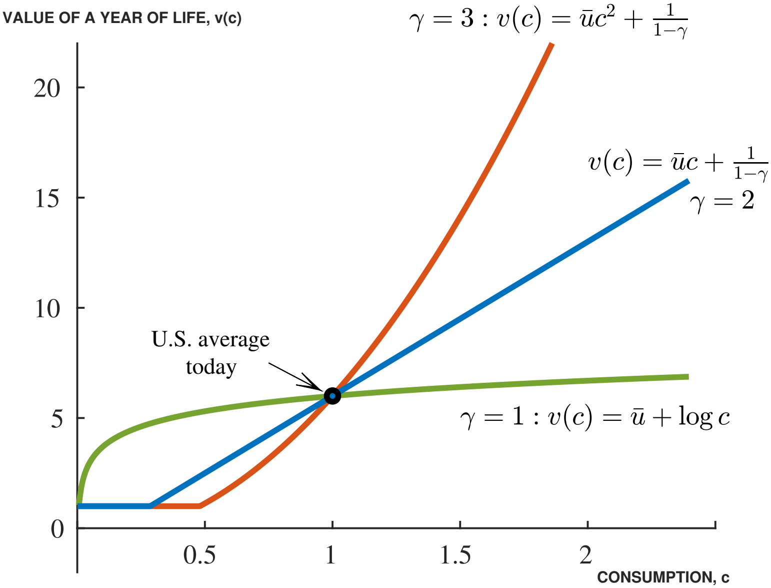
With $\gamma = 2$, dramatically more conservative use of A.I.! Run for 4 years leading to a 57% gain in consumption with a 4% existential risk.

What if $\delta = 2\%$ instead of 1%?

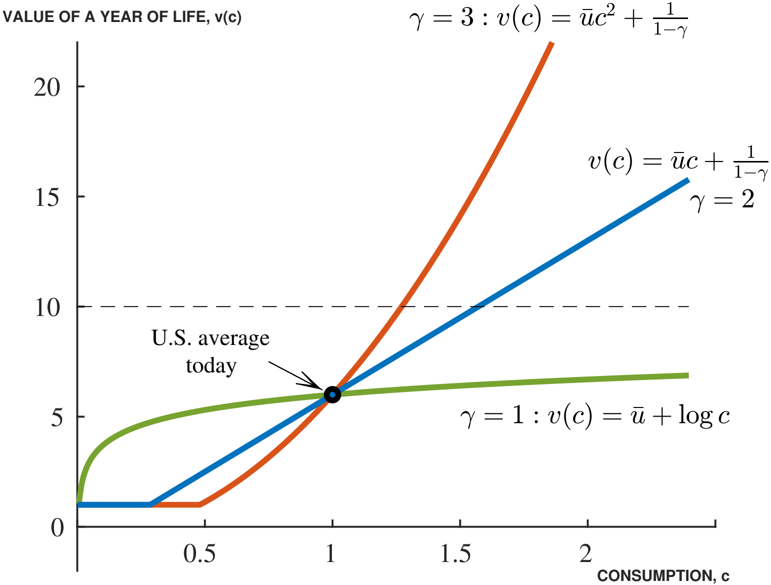
- $g = 10\%$ and $\delta = 2\% \Rightarrow$ AIBC=5 instead of 10.
  - But then $v(c_{us,today}) = 6 > AIBC$

- Therefore it is optimal to set $T^* = 0$ for $\gamma \geq 1$
  - Life is already too valuable relative to the AIBC ratio
  - A.I. is too risky to make even 10% growth worthwhile

Heterogeneity and the Value of Life



Heterogeneity and the Value of Life



Heterogeneity and the Value of Life

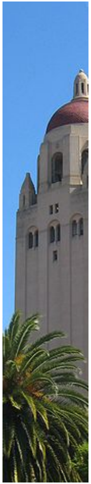
- Heterogeneity by income and consumption
  - Poor people and countries may be much more willing to undertake these risks
  - Rich people less so: income effect in the value of life
- Less risk averse people are more willing to take these risks
  - Including the tech entrepreneurs working on A.I.?

Summary of Simple Model Results

Key Point 1 (Sensitive to δ): Optimal decisions are very sensitive to the magnitude of the A.I. risk. With $\delta = 1\%$ and log utility it is optimal to use the A.I. technology for 40 years involving an overall 1/3 probability of existential risk and a stunning 55-fold increase in consumption. With $\delta = 2\%$, it is optimal to shut it down immediately.

Key Point 2 (Log utility vs CRRA $>$ 1): With $\delta = 1\%$, the optimal decision varies sharply with γ. With $\gamma = 2$, the gain in consumption falls by 100x to 57 percent instead of 55x, the A.I. is used for 4.5 years, and the probability of an existential disaster is just 4 percent.

Decisions are very sensitive to the setup, especially $\gamma = 1$ vs $\gamma \geq 2$



Richer Model

Singularity, improved mortality, and longtermism

Richer Model

- Richer model with dynamics and additional considerations
  - A.I. could lead to a **singularity:** infinite consumption in finite time (immediately)
  - **Mortality improvements:** cure cancer? heart disease?
  - **Longtermism:** what if the social discount rate falls to zero?
  - Adopt A.I. $\Rightarrow$ one-time existential risk probability δ
- No need to calibrate the existential risk. Solve for the **x-risk cutoff** δ^*

$$\delta > \delta^* \Rightarrow \text{ Shut down the A.I.}$$

$$\delta < \delta^* \Rightarrow \text{ Use the A.I.}$$

The Economic Environment

- Social welfare

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

- Lifetime utility

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

  - ○ $N_t = N_0 e^{bt}$ = size of generation t. m = mortality rate.
  - ○ $c_{t+a} = c_0 e^{g(t+a)} = c_t e^{ga}$: identical across people at each date
  - ○ CRRA utility with $\gamma > 1$ here. Set $N_0 = 1$ wlog.

- Should we use the A.I. or not?
  - ○ Shut it down: Growth g_0 and mortality rate m_0
  - ○ Use A.I.: Growth g_{ai} and mortality rate m_{ai}, but one-time existential risk δ

Solution

- Lifetime utility: $$U_t = \frac{\bar{u}}{\rho + m} + \frac{c_t^{1-\gamma}}{1-\gamma} \cdot \frac{1}{\rho + m + (\gamma - 1)g}$$

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

- Use the A.I. as long as $W(g_0, m_0) < (1 - \delta) W(g_{ai}, m_{ai})$, which implies an existential risk cutoff

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

$\delta > \delta^* \Rightarrow$ Shut down the A.I.

$\delta < \delta^* \Rightarrow$ Use the A.I.

Singularity

- What if A.I. results in a Singularity = infinite consumption immediately?

- Key: If $\gamma > 1$, infinite consumption forever delivers finite utility (bounded)

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

- If $m_{ai} = m_0 \equiv m$, then the cutoff is

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

- Comparative statics:
  - δ^*_{sing} falls if $v(c_0), g_0$, or γ is higher
  - δ^*_{sing} rises if ρ_s, ρ, or m is higher (less time for g_0 to kick in)

Baseline parameter values

Discount rates:	$\rho = 1\%$, $\rho_s - b = 1\%$ (e.g. $b = 0$)
Growth without A.I.:	$g_0 = 2\%$
Growth with A.I.:	$g_{ai} = 10\%$
Baseline mortality	$m_0 = 1\%$ (LE=100 years)

Existential Risk Cutoffs: δ^* (no mortality advantage $m_{ai} = m_0$)

γ	$g_{ai} = 10\%$	Singularity
1.01	0.540	0.916
2	0.022	0.024
3	0.005	0.005

- Log utility:
  - High cutoffs confirm Simple Model
  - Singularity $\Rightarrow \delta^* = 1$ for $\gamma \leq 1$

Existential Risk Cutoffs: δ^* (no mortality advantage $m_{ai} = m_0$)

γ	$g_{ai} = 10\%$	Singularity
1.01	0.540	0.916
2	0.022	0.024
3	0.005	0.005

- Log utility:
  - High cutoffs confirm Simple Model
  - Singularity $\Rightarrow \delta^* = 1$ for $\gamma \leq 1$
- CRRA $\gamma \geq 2$:
  - Low cutoffs confirm Simple Model
  - **Singularity similar to $g_{ai} = 10\%$ because flow utility is bounded**

Existential Risk Cutoffs with Improved Mortality: $\delta*$

- What if A.I. cuts mortality in half (doubles life expectancy from 100 to 200 years)?

Existential Risk Cutoffs with Improved Mortality: δ^*

- What if A.I. cuts mortality in half (doubles life expectancy from 100 to 200 years)?

γ	$m_{ai} = m_0 = 1\%$	$m_{ai} = m_0/2 = 0.5\%$
1.01	0.540	0.678
2	0.022	0.267
3	0.005	0.254

- **Answer:** Large increase in the existential risk cutoff!
    - Trading off “lives vs lives” instead of “lives vs consumption”
    - Does not run into the sharp diminishing MU of consumption

Summary

Key Point 3 (Singularities): If $\gamma \leq 1$, the existential risk cutoff for a singularity is $\delta^* = 1$: any risk other than sure annihilation is acceptable to achieve infinite consumption. In contrast, if $\gamma \geq 2$, the cutoffs are much smaller and similar to the cutoffs with $g_{ai} = 10\%$.

Key Point 4 (Mortality improvements): Mortaility risk and existential risk are in the same units and do not run into the diminishing marginal utility of consumption. If A.I. improves life expectancy, the existential risk cutoffs are much higher, on the order of 25–30% for $\gamma = 2$.

Longtermism: What if we put more weight on the future?

- First, with no mortality improvements:

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

  - As $\rho_s \to 0$, then $\delta^*_{sing} \to 0$ (e.g. when $b = 0$)
  - Not worth risking an undiscounted infinite future

Very conservative, as expected

- But what if A.I. also improves mortality?

Longtermism: What if we put more weight on the future? ($g_{ai} = 10\%, b = 0$)

	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

Longtermism: What if we put more weight on the future? ($g_{ai} = 10\%, b = 0$)

	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

- ○ Mortality gains for future generations count more with no discounting

Intuition for near-zero discounting

- As $\rho_s \to 0$, there are essentially infinite generations as rich as Bill Gates
- Therefore each generation has utility

$$U(LE) = \frac{\bar{u}}{\rho + m} = \frac{LE \cdot \bar{u}}{\rho \cdot LE + 1}, \quad \text{where } LE \equiv 1/m = \text{life expectancy}$$

$\rho > 0$ captures diminishing returns to life expectancy

- Can therefore consider a representative generation to get cutoff:

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

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

- With $\rho = 0.01$ and $LE = 100$, we have $\delta^* = .25$

Last Key Point

Key Point 5 (Longtermism): 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, the result is the opposite: putting more weight on future generations means that A.I.-driven mortality improvements are more valuable, making large existential risks worth bearing.

Conclusion: Key Points

- Whether $\gamma = 1$ or $\gamma \geq 2$ matters a lot (bounded utility)
  - With $\gamma \geq 2$, results are often very conservative wrt using A.I.

- Singularities are not so special with bounded utility

- If A.I. improves life expectancy, you are trading off “lives vs lives” and sharply declining MU of consumption is less important $\Rightarrow$ higher cutoffs

- Lower social discounting to put more weight on the future further retains high cutoffs when A.I. improves mortality