Discussion of “How Much Should We Spend to Reduce AI’s Existential Risk?”

Danial Lashkari

Federal Reserve Bank of New York

July 18, 2023

The views expressed here are my own and do not represent those of the Federal Reserve System.

Overview

Admirably bold paper:

Overview

Admirably bold paper:

- An important, hitherto unexplored question (literally a matter of life and death)

Overview

Admirably bold paper:

- An important, hitherto unexplored question (literally a matter of life and death)
- Huge uncertainty about the environment/underlying parameters

Overview

Admirably bold paper:

- An important, hitherto unexplored question (literally a matter of life and death)
- Huge uncertainty about the environment/underlying parameters
- Simple and transparent analysis

Overview

Admirably bold paper:

- An important, hitherto unexplored question (literally a matter of life and death)
- Huge uncertainty about the environment/underlying parameters
- Simple and transparent analysis
- Don’t abandon important problems becasue of lack of data!

Background: AI and Existential Risk

Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.

Signatories:

☑ AI Scientists ☑ Other Notable Figures

Geoffrey Hinton
Emeritus Professor of Computer Science, University of Toronto

Yoshua Bengio
Professor of Computer Science, U. Montreal / Mila

Demis Hassabis
CEO, Google DeepMind

Sam Altman
CEO, OpenAI

Dario Amodei
CEO, Anthropic

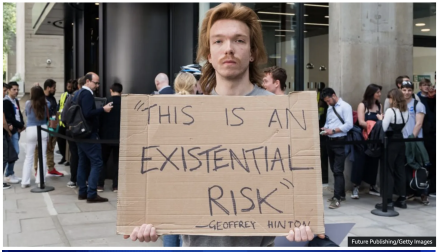
Dawn Song
Professor of Computer Science, UC Berkeley

Ted Lieu
Congressman, US House of Representatives

Bill Gates
Gates Ventures

Ya-Qin Zhang
Professor and Dean, AIR, Tsinghua University

Ilya Sutskever
Co-Founder and Chief Scientist, OpenAI



AI leaders sign statement warning of 'extinction' risk

d

Background: What Are the Existential Risks?

Overview of catastrophic AI risks:

Hendrycks et al. (2023); Bengio et al. (2024)

Malicious Use

- ✗ Bioterrorism
- ✗ Surveillance State
- ✔ Access Restrictions
- ✔ Legal Liability

AI Race

- ✗ Automated Warfare
- ✗ Evolutionary Pressures
- ✔ International Coordination
- ✔ Safety Regulation

Organizational Risks

- ✗ Weak Safety Culture
- ✗ Leaked AI Systems
- ✔ Information Security
- ✔ External Audits

Rogue AIs

- ✗ Power-Seeking
- ✗ Deception
- ✔ Use-Case Restrictions
- ✔ Safety Research

Jonesian Framework for AI X-risk Mitigation

Choose action a affecting two states: Existential Catastrophe (EC) and No Catastrophe (N)

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Jonesian Framework for AI X-risk Mitigation

Choose action a affecting two states: Existential Catastrophe (EC) and No Catastrophe (N)

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_N(a)$$

Jones (2024): action $a \equiv T$ is "intensity of using AI"

$$\pi_{EC}(T) = 1 - e^{-\delta T} \qquad u_{EC}(T) = 0 \qquad u_N(T) = u\left(c_0 e^{gT}\right)$$

Jonesian Framework for AI X-risk Mitigation

Choose action a affecting two states: Existential Catastrophe (EC) and No Catastrophe (N)

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_N(a)$$

Jones (2024): action $a \equiv T$ is "intensity of using AI"

$$\pi_{EC}(T) = 1 - e^{-\delta T} \qquad u_{EC}(T) = 0 \qquad u_N(T) = u\left(c_0 e^{gT}\right)$$

Jones (2025): action $a \equiv x$ is "investment in AI safety"

$$\pi_{EC}(x) = \delta(x) \qquad u_{EC}(x) = u(y - x) \qquad u_N(T) = u(y - x) + \beta V_{t+1}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\left(1 - e^{-\delta T}\right) \times 0 + e^{-\delta T} \times u'\left(c_0 e^{gT}\right) g\, c_0 e^{gT} = u\left(c_0 e^{gT}\right) \times \delta e^{-\delta T}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\frac{u\left(c_0 e^{gT}\right)}{u'\left(c_0 e^{gT}\right)} = \frac{g}{\delta} c_0 e^{gT}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\frac{u\left(c_0 e^{gT}\right)}{u'\left(c_0 e^{gT}\right)} = \frac{g}{\delta} c_0 e^{gT}$$

Refresher on Value of Statistical Life (VSL):

$$\text{const} = (1 - \pi_D)\, u\left(w\left(\pi_D\right)\right) \qquad \Rightarrow \qquad VSL \equiv \frac{\partial w}{\partial \pi_D} = \frac{u}{u'}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\frac{u\left(c_0 e^{gT}\right)}{u'\left(c_0 e^{gT}\right)} = \frac{g}{\delta} c_0 e^{gT}$$

Refresher on Value of Statistical Life (VSL):

$$\text{const} = (1 - \pi_D)\, u\left(w\left(\pi_D\right)\right) \qquad \Rightarrow \qquad VSL \equiv \frac{\partial w}{\partial \pi_D} = \frac{u}{u'} \approx \$250K \approx \boxed{6\, c_0}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\frac{u\left(c_0 e^{gT}\right)}{u'\left(c_0 e^{gT}\right)} = \frac{g}{\delta} c_0 e^{gT} \approx \boxed{10\, c_0 e^{gT}}$$

Refresher on Value of Statistical Life (VSL):

$$\text{const} = (1 - \pi_D)\, u\left(w\left(\pi_D\right)\right) \qquad \Rightarrow \qquad VSL \equiv \frac{\partial w}{\partial \pi_D} = \frac{u}{u'} \approx \$250K \approx \boxed{6\, c_0}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\frac{u\left(c_0 e^{gT}\right)}{u'\left(c_0 e^{gT}\right)} = \frac{g}{\delta} c_0 e^{gT} \approx \boxed{10\, c_0 e^{gT}}$$

Refresher on Value of Statistical Life (VSL):

$$\text{const} = (1 - \pi_D)\, u\left(w\left(\pi_D\right)\right) \qquad \Rightarrow \qquad VSL \equiv \frac{\partial w}{\partial \pi_D} = \frac{u}{u'} \approx \$250K \approx \boxed{6\, c_0}$$

- Income elasticity of VSL:

  Viscusi & Aldy (2003); Costa & Kahn (2004); OECD (2012); Viscusi & Masterman (2017)

$$\frac{d \ln VSL}{d \ln c} = \theta + \frac{1}{VSL/c} \qquad \theta \equiv -\frac{c\, u''(c)}{u'(c)}$$

Jones (2024)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi_{EC}}{\partial a}$$

Jones (2024): $a \equiv T$

$$\frac{u\left(c_0 e^{gT}\right)}{u'\left(c_0 e^{gT}\right)} = \frac{g}{\delta} c_0 e^{gT} \approx \boxed{10\, c_0 e^{gT}}$$

Refresher on Value of Statistical Life (VSL):

$$\text{const} = (1 - \pi_D)\, u\left(w\left(\pi_D\right)\right) \qquad \Rightarrow \qquad VSL \equiv \frac{\partial w}{\partial \pi_D} = \frac{u}{u'} \approx \$250K \approx \boxed{6\, c_0}$$

- Income elasticity of VSL:

  Viscusi & Aldy (2003); Costa & Kahn (2004); OECD (2012); Viscusi & Masterman (2017)

$$\frac{d \ln VSL}{d \ln c} = \theta + \frac{1}{VSL/c} \approx 0.5 - 1 \qquad \theta \equiv -\frac{c\, u''(c)}{u'(c)} \approx 0.3 - 0.8$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$-u'(y - x) = \beta V_+ \times \delta'(x)$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$\frac{1}{VSL} = \frac{u'(c)}{\beta V_+} = -\delta'(x)$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$\frac{1}{\$10M} \approx \frac{u'(c)}{\beta V_+} = -\delta'(x)$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$\frac{x}{\$10M} = \frac{-x\delta'(x)}{\delta(x)} \cdot \delta(x)$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$\frac{x}{\$10M} = \frac{-x\delta'(x)}{\delta(x)} \cdot \delta(x)$$

- Intuitive approach: $\delta \approx 0.01$

$$x \approx \frac{-x\delta'(x)}{\delta(x)} \times \$100K$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$\frac{x}{\$10M} = \frac{-x\delta'(x)}{\delta(x)} \cdot \delta(x)$$

- Intuitive approach: $\delta \approx 0.01$

$$x \approx \frac{-x\delta'(x)}{\delta(x)} \times \$100K \overset{??}{\approx} \boxed{0.01 \times \$100K}$$

Jones (2025)

General model solution:

$$\pi_{EC} \frac{\partial u_{EC}}{\partial a} + (1 - \pi_{EC}) \frac{\partial u_N}{\partial a} = (u_N - u_{EC}) \frac{\partial \pi}{\partial a}$$

Jones (2025): $a \equiv x$

$$\frac{x}{\$10M} = \frac{-x\delta'(x)}{\delta(x)} \cdot \delta(x)$$

- Intuitive approach: $\delta \approx 0.01$

$$x \approx \frac{-x\delta'(x)}{\delta(x)} \times \$100K \overset{??}{\approx} \boxed{0.01 \times \$100K}$$

- Parameterization: $\delta(x) = \delta_o \left(1 - \phi + \phi e^{-\xi^T x}\right)$

Baseline calibation: $x \approx \boxed{0.16 \times \$86K}$

Outline

Framework

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Outline

Framework

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Characterization of...

... the state of Existential Catastrophe (EC)

Outline

Framework

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Characterization of...

... the state of Existential Catastrophe (EC)

... the probability $\pi_{EC}(a)$

Outline

Framework

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Characterization of...

- ... the state of Existential Catastrophe (EC)
- ... the probability $\pi_{EC}(a)$
- ... the choice of action a

Outline

Framework

$$\max_{a} \ \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Characterization of...

- ... the state of Existential Catastrophe (EC)
- ... the probability $\pi_{EC}(a)$
- ... the choice of action a
- ... the objective as a static expected utility maximization problem

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

Parallel with the threat from climate change?

- More predictable trends/probabilities + more concrete catastrophes

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

Parallel with the threat from climate change?

- More predictable trends/probabilities + more concrete catastrophes
- Current share of US spending in mitigation/prevention $\approx 0.6 - 1.2\%$

  Busch & Hsu (2023)

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

Parallel with the threat from climate change?

- More predictable trends/probabilities + more concrete catastrophes
- Current share of US spending in mitigation/prevention $\approx 0.6 - 1.2\%$

  Busch & Hsu (2023)

Existential catastrophes not ending in death?

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

Parallel with the threat from climate change?

- More predictable trends/probabilities + more concrete catastrophes
- Current share of US spending in mitigation/prevention $\approx 0.6 - 1.2\%$

  Busch & Hsu (2023)

Existential catastrophes not ending in death?

- Mass automation-led unemployment?

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

Parallel with the threat from climate change?

- More predictable trends/probabilities + more concrete catastrophes
- Current share of US spending in mitigation/prevention $\approx 0.6 - 1.2\%$

  Busch & Hsu (2023)

Existential catastrophes not ending in death?

- Mass automation-led unemployment?
- Concentration of economic and political power?

The State of AI Catastrophe

Thought-provoking parallel with the pandemic but...

- Covid clearly associated with individual deaths + investment costs well defined
- Existential AI risk more contingent/speculative? Nature of investment less clear?

Parallel with the threat from climate change?

- More predictable trends/probabilities + more concrete catastrophes
- Current share of US spending in mitigation/prevention $\approx 0.6 - 1.2\%$

  Busch & Hsu (2023)

Existential catastrophes not ending in death?

- Mass automation-led unemployment?
- Concentration of economic and political power?
- Loss of intellectual and personal self-actualization through work?

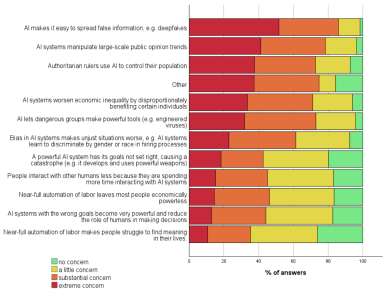
P(doom)

Leading study of forecasts by 2,778 AI experts conducted by Impact AI

Grace et al. (2024)

Likelihood of human extinction

Question	N	Median	Mean
What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species?	1321	5% (IQR 19%)	16.2% (SD 23%)
What probability do you put on human inability to control future advanced AI systems causing human extinction or similarly permanent and severe disempowerment of the human species?	661	10% (IQR 29%)	19.4% (SD 26%)
What probability do you put on future AI advances causing human extinction or similarly permanent and severe disempowerment of the human species within the next 100 years?	655	5% (IQR 19.9%)	14.4% (SD 22.2%)



P(doom)

Leading study of forecasts by 2,778 AI experts conducted by Impact AI

Grace et al. (2024)

Behvioral Biases: probability compression, availability heuristic, overconfidence, framing

P(doom)

Leading study of forecasts by 2,778 AI experts conducted by Impact AI

Grace et al. (2024)

Behvioral Biases: probability compression, availability heuristic, overconfidence, framing

Non-Behvioral Biases?

P(doom)

Leading study of forecasts by 2,778 AI experts conducted by Impact AI

Grace et al. (2024)

Behvioral Biases: probability compression, availability heuristic, overconfidence, framing

Non-Behvioral Biases?

P(doom)

Leading study of forecasts by 2,778 AI experts conducted by Impact AI

Grace et al. (2024)

Behvioral Biases: probability compression, availability heuristic, overconfidence, framing

Non-Behvioral Biases?

> *"What they are doing is running a well-funded panic campaign. [...] A better representation of this survey would indicate that it was funded, phrased, and analyzed by 'x-risk' effective altruists. Behind 'AI Impacts' and other 'AI Safety' organizations, there's a well-oiled 'x-risk' machine. When the media is covering them, it has to mention it."*
>
> *Weiss-Blatt (author/researcher)*

P(doom)

Leading study of forecasts by 2,778 AI experts conducted by Impact AI

Grace et al. (2024)

Behvioral Biases: probability compression, availability heuristic, overconfidence, framing

Non-Behvioral Biases?

> *"As in previous years, many of the questions are asked from the AI-doomer, existential-risk perspective. [...] I still think the focus is on 'How much should we worry?' rather than on doing a careful risk analysis and setting policy to mitigate the relevant risks."*
>
> *Dietterich (former president of AAAI)*

Our Choice Space?

Distinguishing intensity of AI development (T) from investments in EC mitigation (x)?

Our Choice Space?

Distinguishing intensity of AI development (T) from investments in EC mitigation (x)?

- Most mitigation proposals in Bengio et al. (2024) have implications for AI development

Our Choice Space?

Distinguishing intensity of AI development (T) from investments in EC mitigation (x)?

- Most mitigation proposals in Bengio et al. (2024) have implications for AI development
- Context: current share of AI investments in US GDP $\approx 1\%$

Our Choice Space?

Distinguishing intensity of AI development (T) from investments in EC mitigation (x)?

- Most mitigation proposals in Bengio et al. (2024) have implications for AI development
- Context: current share of AI investments in US GDP $\approx 1\%$

Are the private decisions inefficient? What are the externalities/spillovers?

Framework: Static Expected Utility Maximization

Individual vs. Societal Objective Function:

Framework: Static Expected Utility Maximization

Individual vs. Societal Objective Function:

- Human extinction $\stackrel{?}{=}$ Death of all individuals

Framework: Static Expected Utility Maximization

Individual vs. Societal Objective Function:

- Human extinction $\stackrel{?}{=}$ Death of all individuals

Risk vs. Uncertainty:

- Ambiguity aversion?

$$\max_{a} \min_{,\pi_{EC}(a)\in\Pi} \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_{N}(a)$$

Framework: Static Expected Utility Maximization

Individual vs. Societal Objective Function:

- Human extinction $\stackrel{?}{=}$ Death of all individuals

Risk vs. Uncertainty:

- Ambiguity aversion?

$$\max_{a} \min_{,\pi_{EC}(a)\in\Pi} \pi_{EC}(a)\ u_{EC}(a) + (1 - \pi_{EC}(a))\ u_N(a)$$

Dynamics:

Framework: Static Expected Utility Maximization

Individual vs. Societal Objective Function:

- Human extinction $\stackrel{?}{=}$ Death of all individuals

Risk vs. Uncertainty:

- Ambiguity aversion?

$$\max_{a} \min_{,\pi_{EC}(a)\in\Pi} \pi_{EC}(a)\ u_{EC}(a) + (1-\pi_{EC}(a))\ u_N(a)$$

Dynamics:

- The decision to adopt and information about likely disaster jointly unfold over time

  Acemoglu & Lensman (2024)

$$V(z_t, s_t) = \max_{x,T} [1-\pi_{EC}(x, z_t, s_t)]\ (u(z_t - T - x) + \beta\, \mathbb{E}[V(z_{t+1}, s_{t+1}) | T, x])$$

Conclusion

Admirably bold paper on important question: how much should we spend to avoid AI doom?

Simple framework + transparent assumptions $\Rightarrow$ Provocative answer

Starting point for much future work...

Conclusion

Admirably bold paper on important question: how much should we spend to avoid AI doom?

Simple framework + transparent assumptions $\Rightarrow$ Provocative answer

Starting point for much future work...

Don't abandon important problems becasue of lack of data!