AI Policy

Miles Brundage

AGENDA

01 BACKGROUND

- Policy in general
- My road to AI and AI policy
- Some things I worked on at OpenAI

02 AI POLICY IN GENERAL

- Key concepts
- Key tensions

03 SOME HOT TAKES

- [redacted - you'll have to wait until later]

04 WHERE YOU FIT IN

- Policy-related research/engineering opportunities
- Your voice in companies and public discussions

01

Background

“Policy”

“X policy” basically just means “the decisions that society makes about X, and how they are and should be made.”

AI policy/governance is the theory and practice of governmental and non-governmental decision-making about AI.

“Policy”

Healthcare policy: how insurance is regulated, how drug approvals work, etc.

Energy policy: how utilities are regulated, how research and development is encouraged, etc.

We’re still figuring out exactly what AI policy should involve.

“Policy”

Regulation is a part of it, but not all of it.

E.g. the CHIPS Act (subsidizing the American semiconductor industry) is an example of AI policy.

It’s also not all about the government, either.

“Policy”

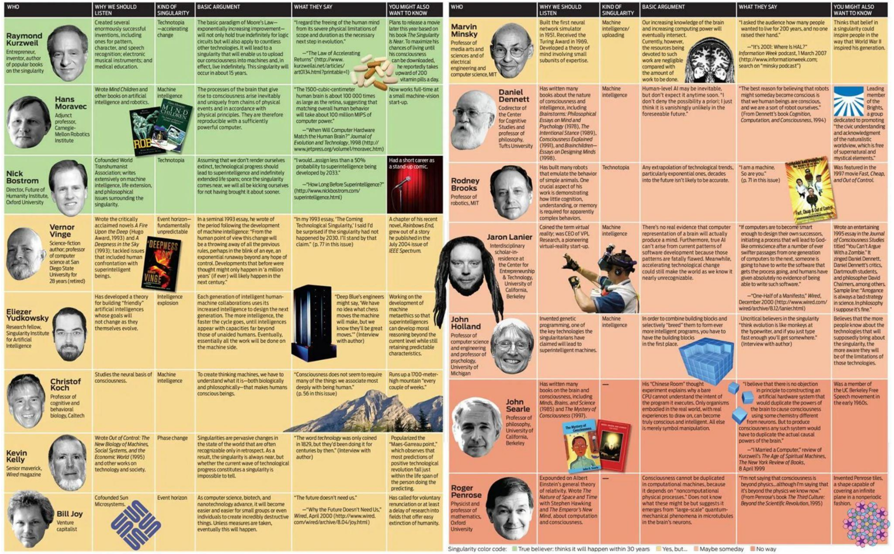
- A lot of AI policy is informed by and related to safety, broadly defined, including alignment, reliability, etc.
- But there’s more to AI policy than safety, and I won’t say much about safety on its own here

AI policy examples

Illustratively, AI policy involves:

- Company decision-making about the right use case policies for APIs and first party products
- Company decisions around deployment of technologies
- Regulations like the EU AI Act
- Industry norms/best practices
- Academics' decisions about open sourcing and documenting models
- Etc.

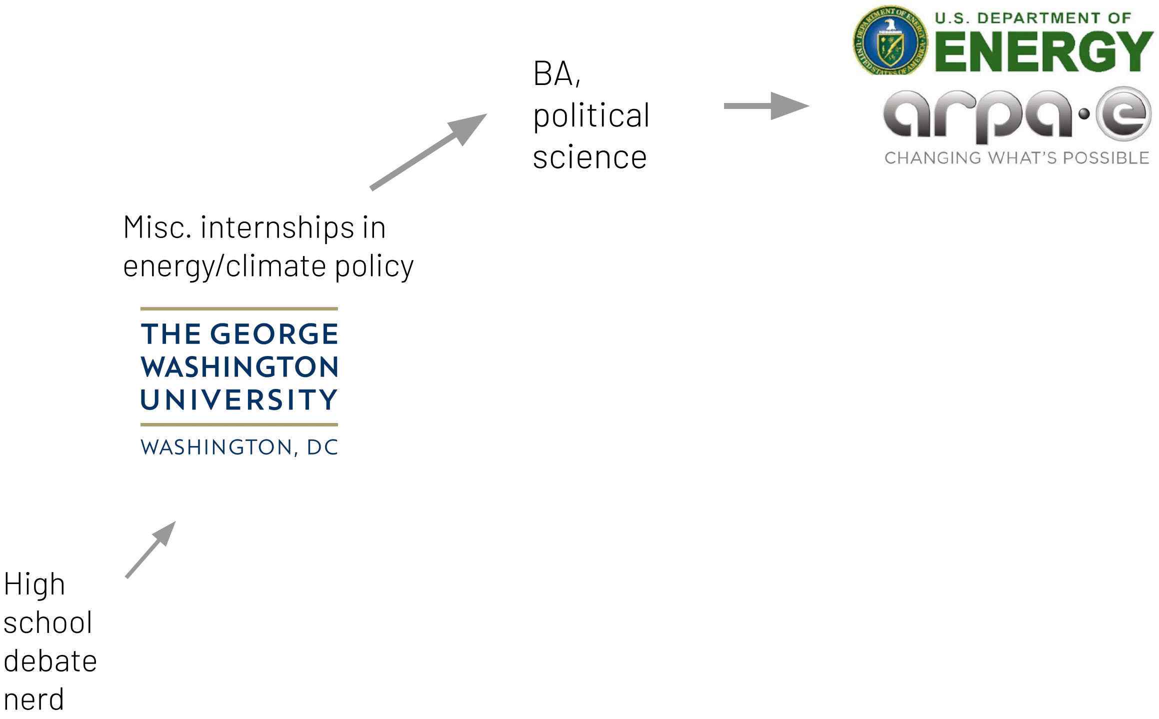
My career arc



WHO	WHY WE SHOULD LISTEN	KIND OF SINGULARITY	BASIC ARGUMENT	WHAT THEY SAY	YOU MIGHT ALSO WANT TO KNOW
Raymond Kurzweil Entrepreneur, inventor, author of popular books on the singularity	Created several enormously successful inventions, including ones for pattern, character, and speech recognition; electronic musical instruments; and medical education.	Technotopia —accelerating change	The basic paradigm of Moore's Law—exponentially increasing improvement—will not only hold true indefinitely for logic circuits but will also apply to countless other technologies. It will lead to a singularity that will enable us to upload our consciousness into machines and, in effect, live indefinitely. This singularity will occur in about 15 years.	"I regard the freeing of the human mind from its severe physical limitations of scope and duration as the necessary next step in evolution." —"The Law of Accelerating Returns" (http://www.kurzweilai.net/articles/art0134.html?printable=1)	Plans to release a movie later this year based on his book *The Singularity Is Near*. To maximize his chances of living until his consciousness can be downloaded, he reportedly takes upward of 200 vitamin pills a day.
Hans Moravec Adjunct professor, Carnegie-Mellon Robotics Institute	Wrote *Mind Children* and other books on artificial intelligence and robotics.	Machine intelligence	The processes of the brain that give rise to consciousness arise inevitably and uniquely from chains of physical events and in accordance with physical principles. They are therefore reproducible with a sufficiently powerful computer.	"The 1500-cubic-centimeter human brain is about 100 000 times as large as the retina, suggesting that matching overall human behavior will take about 100 million MIPS of computer power." —"When Will Computer Hardware Match the Human Brain?" *Journal of Evolution and Technology*, 1998 (http://www.jetpress.org/volume1/moravec.htm)	Now works full-time at a small machine-vision start-up.
Nick Bostrom Director, Future of Humanity Institute, Oxford University	Cofounded World Transhumanist Association; writes extensively on machine intelligence, life extension, and philosophical issues surrounding the singularity.	Technotopia	Assuming that we don't render ourselves extinct, technological progress should lead to superintelligence and indefinitely extended life spans; once the singularity comes near, we will all be kicking ourselves for not having brought it about sooner.	"I would...assign less than a 50% probability to superintelligence being developed by 2033." —"How Long Before Superintelligence?" (http://www.nickbostrom.com/superintelligence.html)	Had a short career as a stand-up comic.
Vernor Vinge Science-fiction author; professor of computer science at San Diego State University for 28 years (retired)	Wrote the critically acclaimed novels *A Fire Upon the Deep* (Hugo Award, 1993) and *A Deepness in the Sky* (1993); tackled issues that included human confrontation with superintelligent beings.	Event horizon—fundamentally unpredictable	In a seminal 1993 essay, he wrote of the period following the development of machine intelligence: "From the human point of view this change will be a throwing away of all the previous rules, perhaps in the blink of an eye, an exponential runaway beyond any hope of control. Developments that before were thought might only happen in 'a million years' (if ever) will likely happen in the next century."	"In my 1993 essay, 'The Coming Technological Singularity,' I said I'd be surprised if the singularity had not happened by 2030. I'll stand by that claim." (p. 77 in this issue)	A chapter of his recent novel, *Rainbows End*, grew out of a story he published in the July 2004 issue of *IEEE Spectrum*.
Eliezer Yudkowsky Research fellow, Singularity Institute for Artificial Intelligence	Has developed a theory for building "friendly" artificial intelligences whose goals will not change as they themselves evolve.	Intelligence explosion	Each generation of intelligent human-machine collaborations uses its increased intelligence to design the next generation. The more intelligence, the faster the cycle goes, until intelligences appear with capacities far beyond those of unaided humans. Eventually, essentially all the work will be done on the machine side.	"Deep Blue's engineers might say, 'We have no idea what chess moves the machine will make, but we know they'll be great moves.'" (Interview with author)	Working on the development of machine metaethics so that superintelligences can develop moral reasoning beyond the current level while still retaining predictable characteristics.
Christof Koch Professor of cognitive and behavioral biology, Caltech	Studies the neural basis of consciousness.	Machine intelligence	To create thinking machines, we have to understand what it is—both biologically and philosophically—that makes humans conscious beings.	"Consciousness does not seem to require many of the things we associate most deeply with being human." (p. 56 in this issue)	Runs up a 1700-meter-high mountain "every couple of weeks."
Kevin Kelly Senior maverick, *Wired* magazine	Wrote *Out of Control: The New Biology of Machines, Social Systems, and the Economic World* (1995) and other works on technology and society.	Phase change	Singularities are pervasive changes in the state of the world that are often recognizable only in retrospect. As a result, the singularity is always near, but whether the current wave of technological progress constitutes a singularity is impossible to tell.	"The word *technology* was only coined in 1829, but they'd been doing it for centuries by then." (Interview with author)	Popularized the "Maes-Garreau point," which observes that most predictions of positive technological revolution fall just within the life span of the person doing the predicting.
Bill Joy Venture capitalist	Cofounded Sun Microsystems.	Event horizon	As computer science, biotech, and nanotechnology advance, it will become easier and easier for small groups or even individuals to create incredibly destructive things. Unless measures are taken, eventually this will happen.	"The future doesn't need us." —"Why the Future Doesn't Need Us," *Wired*, April 2000 (http://www.wired.com/wired/archive/8.04/joy.html)	Has called for voluntary renunciation or at least a delay of research into fields that offer easy extinction of humanity.
Marvin Minsky Professor of media arts and sciences and of electrical engineering and computer science, MIT	Built the first neural network simulator in 1951. Received the Turing Award in 1969. Developed a theory of mind involving small subunits of expertise.	Machine intelligence/ uploading	Our increasing knowledge of the brain and increasing computing power will eventually intersect. Currently, however, the resources being devoted to such work are negligible compared with the amount of work to be done.	"I asked the audience how many people wanted to live for 200 years, and no one raised their hand." —"It's 2001: Where Is HAL?" *Information Week* podcast, 1 March 2007 (http://www.informationweek.com; search on "minsky podcast")	Thinks that belief in a singularity could inspire people in the way that World War II inspired his generation.
Daniel Dennett Codirector of the Center for Cognitive Studies and professor of philosophy, Tufts University	Has written many books about the nature of consciousness and intelligence, including *Brainstorms: Philosophical Essays on Mind and Psychology* (1978), *The Intentional Stance* (1989), *Consciousness Explained* (1991), and *Brainchildren—Essays on Designing Minds* (1998).	Machine intelligence	Human-level AI may be inevitable, but don't expect it anytime soon. "I don't deny the possibility a priori; I just think it is vanishingly unlikely in the foreseeable future."	"The best reason for believing that robots might someday become conscious is that we human beings are conscious, and we are a sort of robot ourselves." (From Dennett's book *Cognition, Computation, and Consciousness*, 1994)	Leading member of the Brights, a group dedicated to promoting "the civic understanding and acknowledgment of the naturalistic worldview, which is free of supernatural and mystical elements."
Rodney Brooks Professor of robotics, MIT	Has built many robots that emulate the behavior of simple animals. One crucial aspect of his work is demonstrating how little cognition, understanding, or memory is required for apparently complex behaviors.	Technotopia	Any extrapolation of technological trends, particularly exponential ones, decades into the future isn't likely to be accurate.	"I am a machine. So are you." (p. 71 in this issue)	Was featured in the 1997 movie *Fast, Cheap, and Out of Control*.
Jaron Lanier Interdisciplinary scholar-in-residence at the Center for Entrepreneurship & Technology, University of California, Berkeley	Coined the term *virtual reality*; was CEO of VPL Research, a pioneering virtual-reality start-up.	Machine intelligence	There's no real evidence that computer representation of a brain will actually produce a mind. Furthermore, true AI can't arise from current patterns of software development because those patterns are fatally flawed. Meanwhile, accelerating technological change could still make the world as we know it nearly unrecognizable.	"If computers are to become smart enough to design their own successors, initiating a process that will lead to God-like omniscience after a number of ever swifter passages from one generation of computers to the next, someone is going to have to write the software that gets the process going, and humans have given absolutely no evidence of being able to write such software." —"One-Half of a Manifesto," *Wired*, December 2000 (http://www.wired.com/wired/archive/8.12/lanier.html)	Wrote an entertaining 1995 essay in the *Journal of Consciousness Studies* titled "You Can't Argue With a Zombie." It zinged Daniel Dennett, Daniel Dennett's critics, Dartmouth students, and philosopher David Chalmers, among others. Sample line: "Arrogance is always a bad strategy in science. In philosophy I suppose it's fine."
John Holland Professor of computer science and engineering and professor of psychology, University of Michigan	Invented genetic programming, one of the key technologies the singularitarians have claimed will lead to superintelligent machines.	Machine intelligence	In order to combine building blocks and selectively "breed" them to form ever more intelligent programs, you have to have the building blocks in the first place.	Uncritical believers in the singularity "think evolution is like monkeys at the typewriter, and if you just type fast enough you'll get somewhere." (Interview with author)	Believes that the more people know about the technologies that will supposedly bring about the singularity, the more aware they will be of the limitations of those technologies.
John Searle Professor of philosophy, University of California, Berkeley	Has written many books on the brain and consciousness, including *Minds, Brains, and Science* (1985) and *The Mystery of Consciousness* (1997).	—	His "Chinese Room" thought experiment explains why a bare CPU cannot understand the intent of the program it executes. Only organisms embodied in the real world, with real experiences to draw on, can become truly conscious and intelligent. All else is merely symbol manipulation.	"I believe that there is no objection in principle to constructing an artificial hardware system that would duplicate the powers of the brain to cause consciousness using some chemistry different from neurons. But to produce consciousness any such system would have to duplicate the actual causal powers of the brain." —"I Married a Computer," review of Kurzweil's *The Age of Spiritual Machines*, *The New York Review of Books*, 8 April 1999	Was a member of the UC Berkeley Free Speech movement in the early 1960s.
Roger Penrose Physicist and professor of mathematics, Oxford University	Expounded on Albert Einstein's general theory of relativity. Wrote *The Nature of Space and Time* with Stephen Hawking and *The Emperor's New Mind*, about computation and consciousness.	—	Consciousness cannot be duplicated in computational machines, because it depends on "noncomputational physical processes." Does not know what these might be but suggests it emerges from "large-scale" quantum-mechanical phenomena in microtubules in the brain's neurons.	"I'm not saying that consciousness is beyond physics...although I'm saying that it's beyond the physics we know now." (From Penrose's book *The Third Culture: Beyond the Scientific Revolution*, 1995)	Invented Penrose tiles, a shape capable of covering an infinite plane in a nonperiodic fashion.

Singularity color code: 🟩 True believer: thinks it will happen within 30 years 🟨 Yes, but... 🟧 Maybe someday 🟥 No way

IEEE Spectrum, 2008



High school debate nerd

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Misc. internships in energy/climate policy

BA, political science

U.S. DEPARTMENT OF ENERGY

arpa·e

CHANGING WHAT'S POSSIBLE

My career arc

My career arc

High school debate nerd

Misc. internships in energy/climate policy

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

BA, political science

U.S. DEPARTMENT OF ENERGY

arpa·e
CHANGING WHAT'S POSSIBLE

ASU

No AI policy grants :(

Image credit: Britannica

Still took some AI classes

My career arc

High school debate nerd

THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Misc. internships in energy/climate policy

BA, political science

U.S. DEPARTMENT OF ENERGY
arpa·e
CHANGING WHAT'S POSSIBLE

ASU

No AI policy grants :(

Image credit: Britannica

Still took some AI classes

PhD almost done

UNIVERSITY OF OXFORD

AI policy grant $ in 2015 (via Elon Musk)

My career arc

High school debate nerd

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Misc. internships in energy/climate policy

BA, political science

U.S. DEPARTMENT OF ENERGY

arpa·e

CHANGING WHAT'S POSSIBLE

ASU

No AI policy grants :(

Image credit: Britannica

Still took some AI classes

PhD almost done

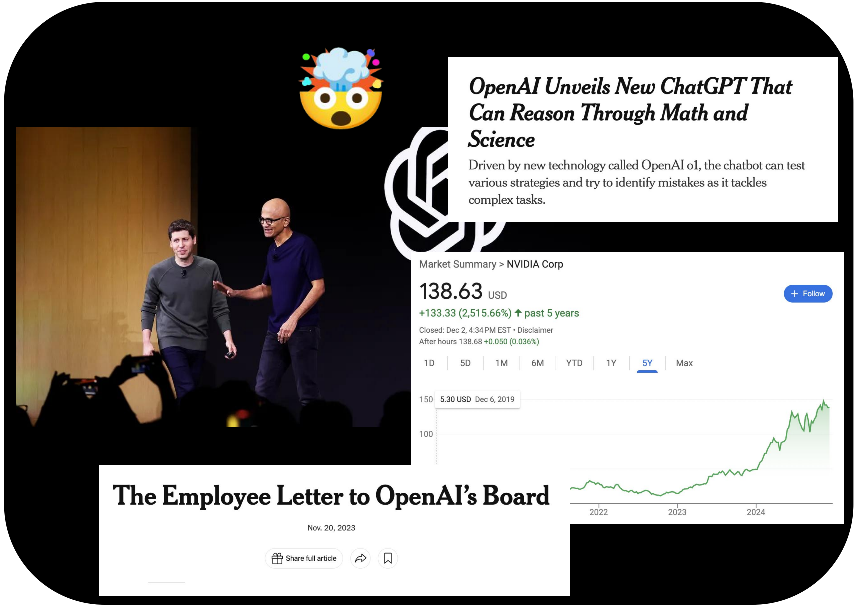
UNIVERSITY OF OXFORD

AI policy grant $ in 2015 (via Elon Musk)

Start collaborating with OpenAI folks

OpenAI

My time at OpenAI in short



OpenAI Unveils New ChatGPT That Can Reason Through Math and Science

Driven by new technology called OpenAI o1, the chatbot can test various strategies and try to identify mistakes as it tackles complex tasks.

Market Summary > NVIDIA Corp

138.63 USD

+133.33 (2,515.66%) ↑ past 5 years

Closed: Dec 2, 4:34 PM EST • Disclaimer

After hours 138.68 +0.050 (0.036%)

The Employee Letter to OpenAI's Board

Nov. 20, 2023

Not really - GPT-2→3, Dota, etc. were quite exciting

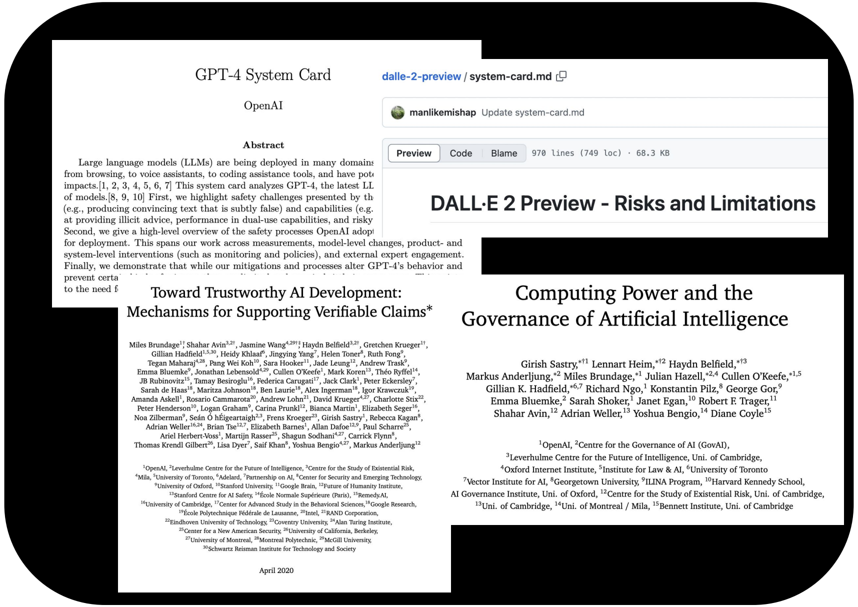
A few things I worked on

Analysis of and policies for, e.g.:

- GPT-2, 3, 3.5, **4**, 4v, 4o
- **Codex**
- DALL-E 2, 3
- CLIP
- **o1**

Helping shape/scale red teaming, economic impact analysis, etc.

More general research on, e.g., agents, compute, frontier AI regulation, etc.



GPT-4 System Card

OpenAI

Abstract

Large language models (LLMs) are being deployed in many domains
from browsing, to voice assistants, to coding assistance tools, and have pote
impacts.[1, 2, 3, 4, 5, 6, 7] This system card analyzes GPT-4, the latest LL
of models.[8, 9, 10] First, we highlight safety challenges presented by th
(e.g., producing convincing text that is subtly false) and capabilities (e.g.
at providing illicit advice, performance in dual-use capabilities, and risky
Second, we give a high-level overview of the safety processes OpenAI adopt
for deployment. This spans our work across measurements, model-level changes, product- and system-level interventions (such as monitoring and policies), and external expert engagement. Finally, we demonstrate that while our mitigations and processes alter GPT-4's behavior and
prevent certa
to the need f

dalle-2-preview / system-card.md

manlikemishap Update system-card.md

970 lines (749 loc) · 68.3 KB

DALL·E 2 Preview - Risks and Limitations

Toward Trustworthy AI Development:
Mechanisms for Supporting Verifiable Claims*

Miles Brundage[1†], Shahar Avin[3,2†], Jasmine Wang[4,29†‡], Haydn Belfield[3,2†], Gretchen Krueger[1†], Gillian Hadfield[1,5,30], Heidy Khlaaf[6], Jingying Yang[7], Helen Toner[8], Ruth Fong[9], Tegan Maharaj[4,28], Pang Wei Koh[10], Sara Hooker[11], Jade Leung[12], Andrew Trask[9], Emma Bluemke[9], Jonathan Lebensold[4,29], Cullen O'Keefe[1], Mark Koren[13], Théo Ryffel[14], JB Rubinovitz[15], Tamay Besiroglu[16], Federica Carugati[17], Jack Clark[1], Peter Eckersley[7], Sarah de Haas[18], Maritza Johnson[18], Ben Laurie[18], Alex Ingerman[18], Igor Krawczuk[19], Amanda Askell[1], Rosario Cammarota[20], Andrew Lohn[21], David Krueger[4,27], Charlotte Stix[22], Peter Henderson[10], Logan Graham[9], Carina Prunkl[12], Bianca Martin[1], Elizabeth Seger[16], Noa Zilberman[9], Seán Ó hÉigeartaigh[2,3], Frens Kroeger[23], Girish Sastry[1], Rebecca Kagan[8], Adrian Weller[16,24], Brian Tse[12,7], Elizabeth Barnes[1], Allan Dafoe[12,9], Paul Scharre[25], Ariel Herbert-Voss[1], Martijn Rasser[25], Shagun Sodhani[4,27], Carrick Flynn[8], Thomas Krendl Gilbert[26], Lisa Dyer[7], Saif Khan[8], Yoshua Bengio[4,27], Markus Anderljung[12]

[1]OpenAI, [2]Leverhulme Centre for the Future of Intelligence, [3]Centre for the Study of Existential Risk, [4]Mila, [5]University of Toronto, [6]Adelard, [7]Partnership on AI, [8]Center for Security and Emerging Technology, [9]University of Oxford, [10]Stanford University, [11]Google Brain, [12]Future of Humanity Institute, [13]Stanford Centre for AI Safety, [14]École Normale Supérieure (Paris), [15]Remedy.AI, [16]University of Cambridge, [17]Center for Advanced Study in the Behavioral Sciences, [18]Google Research, [19]École Polytechnique Fédérale de Lausanne, [20]Intel, [21]RAND Corporation, [22]Eindhoven University of Technology, [23]Coventry University, [24]Alan Turing Institute, [25]Center for a New American Security, [26]University of California, Berkeley, [27]University of Montreal, [28]Montreal Polytechnic, [29]McGill University, [30]Schwartz Reisman Institute for Technology and Society

April 2020

Computing Power and the
Governance of Artificial Intelligence

Girish Sastry,[*†1] Lennart Heim,[*†2] Haydn Belfield,[*†3] Markus Anderljung,[*2] Miles Brundage,[*1] Julian Hazell,[*2,4] Cullen O'Keefe,[*1,5] Gillian K. Hadfield,[*6,7] Richard Ngo,[1] Konstantin Pilz,[8] George Gor,[9] Emma Bluemke,[2] Sarah Shoker,[1] Janet Egan,[10] Robert F. Trager,[11] Shahar Avin,[12] Adrian Weller,[13] Yoshua Bengio,[14] Diane Coyle[15]

[1]OpenAI, [2]Centre for the Governance of AI (GovAI), [3]Leverhulme Centre for the Future of Intelligence, Uni. of Cambridge, [4]Oxford Internet Institute, [5]Institute for Law & AI, [6]University of Toronto [7]Vector Institute for AI, [8]Georgetown University, [9]ILINA Program, [10]Harvard Kennedy School, AI Governance Institute, Uni. of Oxford, [12]Centre for the Study of Existential Risk, Uni. of Cambridge, [13]Uni. of Cambridge, [14]Uni. of Montreal / Mila, [15]Bennett Institute, Uni. of Cambridge

02

AI Policy in General

Key concepts in AI policy

I’ll try to be mostly* uncontroversial in this section, compared to the next one

*there is basically nothing that’s totally uncontroversial in AI policy

AI as a general-purpose technology

Images: IEA, Britannica

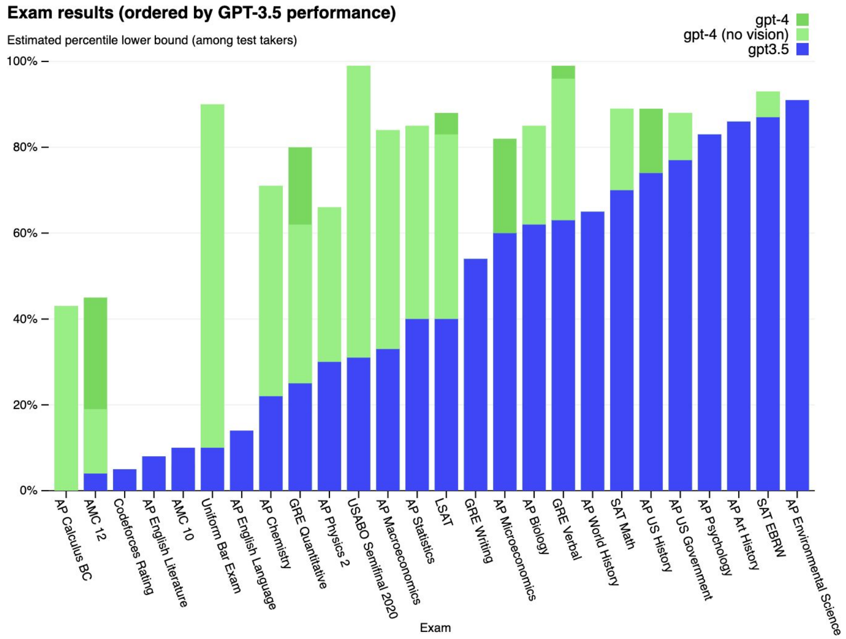
AI as a general-purpose technology

→ will have impacts across all sectors

→ "AI policy" will become, or interact with, "everything policy" by default (how to avoid overreach?)

→ can be differentially sped up/slow down in some aspects, though such interventions will degrade by default

AI is a fast-moving technology

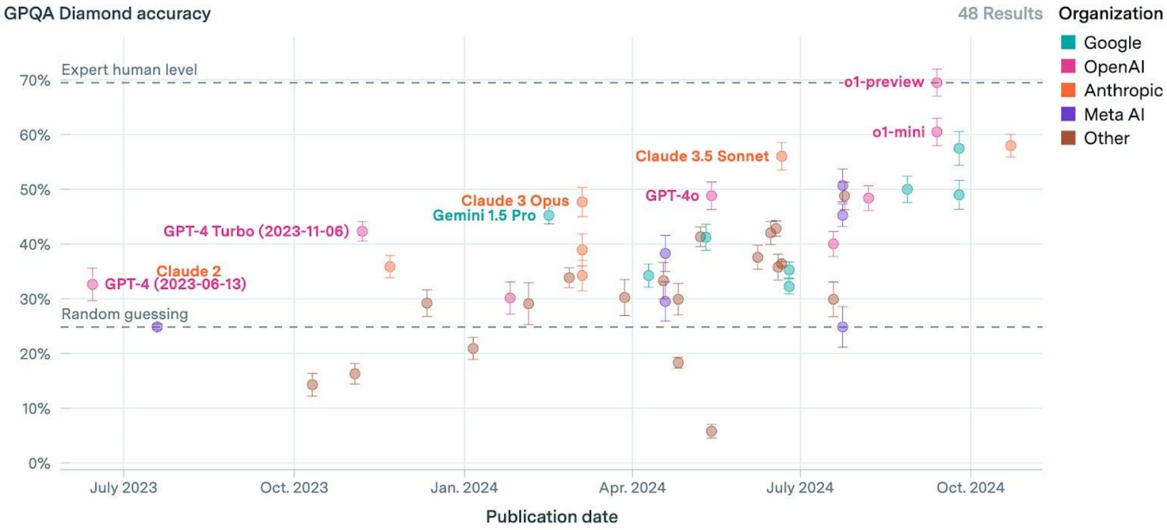


OpenAI

AI is a fast-moving technology

AI performance on a set of Ph.D.-level science questions

EPOCH AI



CC-BY

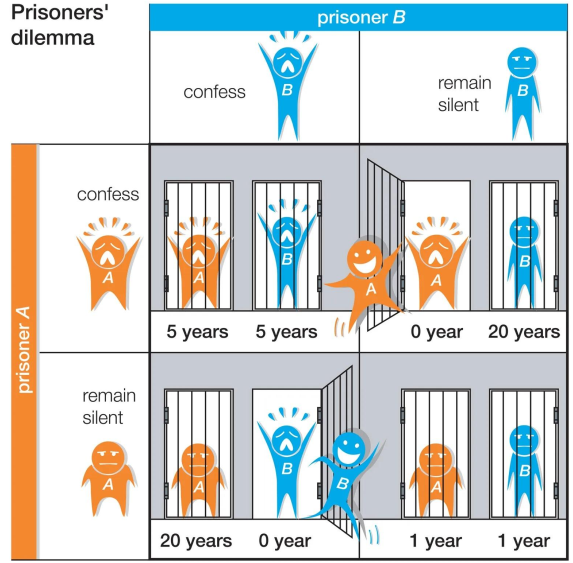
epoch.ai

AI is a fast-moving technology

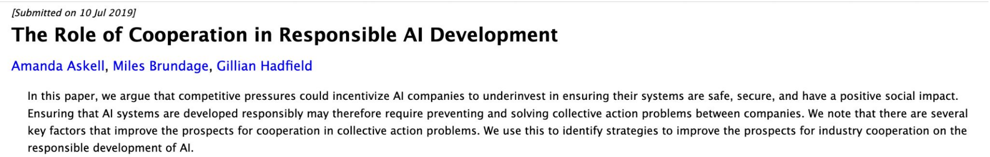
It also features semi-regular paradigm shifts:

- RL in videogames/simulations →
- Large-scale unsupervised pretraining + a bit of SL supervised and reinforcement learning at the end →
- “Actual RL” on language models

AI development is (in part) a collective action problem



© 2010 Encyclopædia Britannica, Inc.

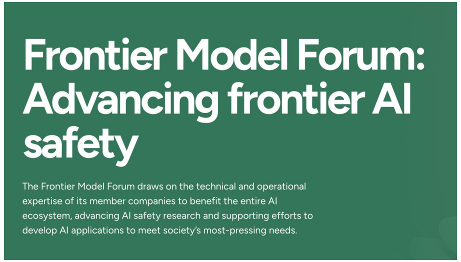


[Submitted on 10 Jul 2019]

The Role of Cooperation in Responsible AI Development

Amanda Askell, Miles Brundage, Gillian Hadfield

In this paper, we argue that competitive pressures could incentivize AI companies to underinvest in ensuring their systems are safe, secure, and have a positive social impact. Ensuring that AI systems are developed responsibly may therefore require preventing and solving collective action problems between companies. We note that there are several key factors that improve the prospects for cooperation in collective action problems. We use this to identify strategies to improve the prospects for industry cooperation on the responsible development of AI.



Frontier Model Forum: Advancing frontier AI safety

The Frontier Model Forum draws on the technical and operational expertise of its member companies to benefit the entire AI ecosystem, advancing AI safety research and supporting efforts to develop AI applications to meet society's most-pressing needs.

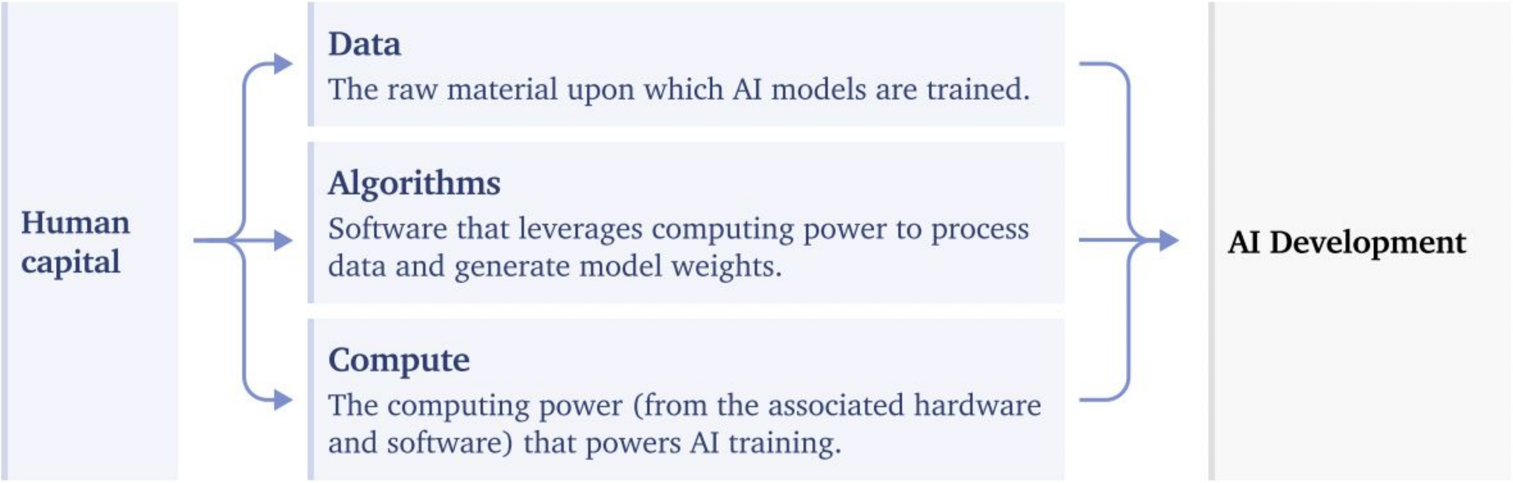
SEPTEMBER 12, 2023

FACT SHEET: Biden-Harris Administration Secures Voluntary Commitments from Eight Additional Artificial Intelligence Companies to Manage the Risks Posed by AI

BRIEFING ROOM › STATEMENTS AND RELEASES

Not nec. this exact dilemma

Different inputs to AI capabilities are easier/harder to govern than others



Human capital

Data

The raw material upon which AI models are trained.

Algorithms

Software that leverages computing power to process data and generate model weights.

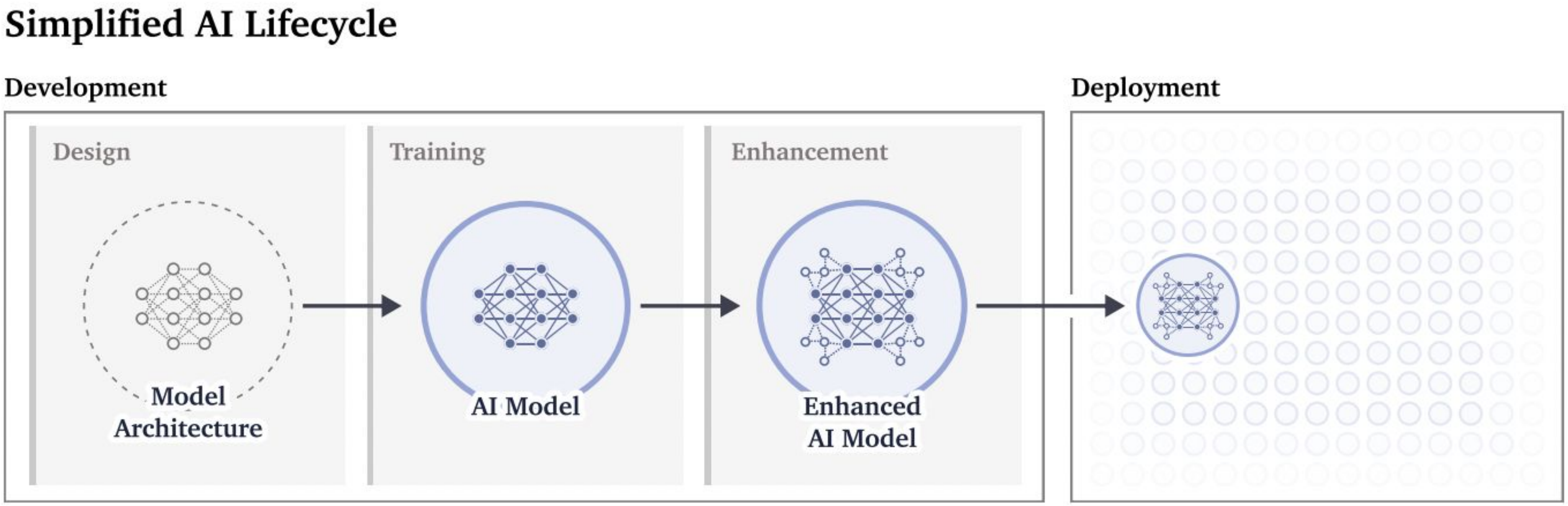
Compute

The computing power (from the associated hardware and software) that powers AI training.

AI Development

Sastry et al., 2024

Different phases of AI development and deployment raise different policy questions



Copyright?

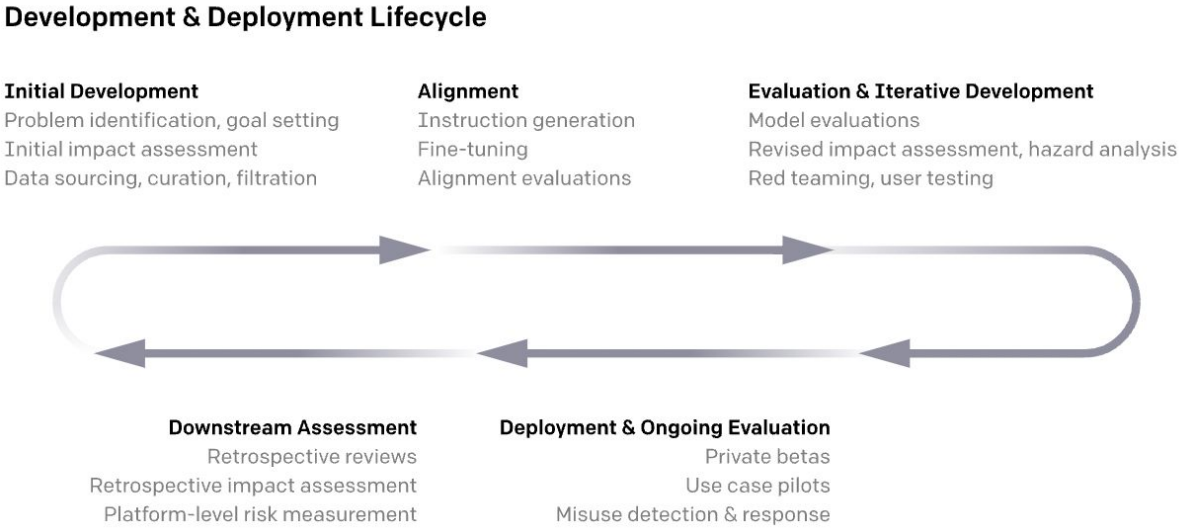
Bias?

Misuse?

Economic impact?

Sastry et al., 2024

Different phases of AI development and deployment raise different policy questions



Development & Deployment Lifecycle

Initial Development
Problem identification, goal setting
Initial impact assessment
Data sourcing, curation, filtration

Alignment
Instruction generation
Fine-tuning
Alignment evaluations

Evaluation & Iterative Development
Model evaluations
Revised impact assessment, hazard analysis
Red teaming, user testing

Deployment & Ongoing Evaluation
Private betas
Use case pilots
Misuse detection & response

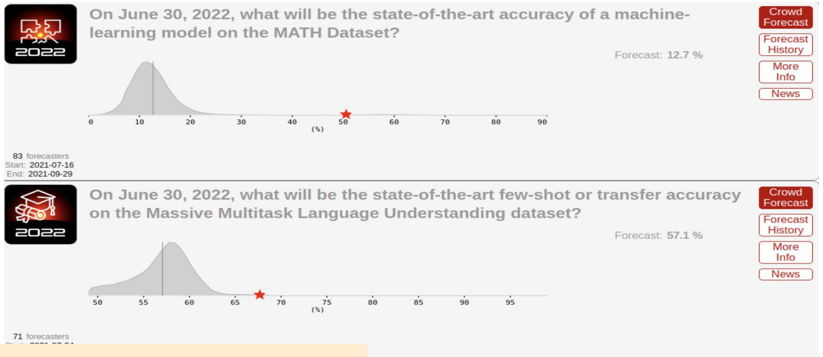
Downstream Assessment
Retrospective reviews
Retrospective impact assessment
Platform-level risk measurement

OpenAI

AI is still behind many other areas of policy, analytically

Many AI forecasts/opinions etc. are way off, or unfalsifiable/“not even wrong”

Many seemingly big topics are basically brand new (e.g. test-time compute)



On June 30, 2022, what will be the state-of-the-art accuracy of a machine-learning model on the MATH Dataset?

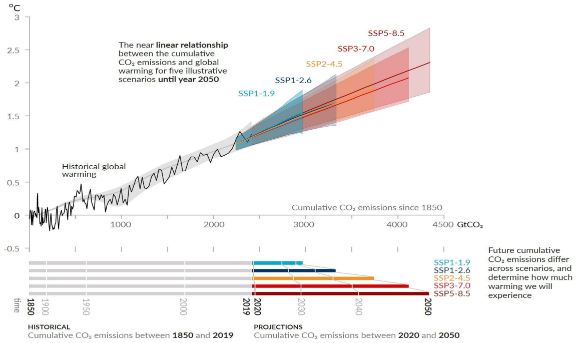
Forecast: 12.7 %

83 forecasters
Start: 2021-07-16
End: 2021-09-29

On June 30, 2022, what will be the state-of-the-art few-shot or transfer accuracy on the Massive Multitask Language Understanding dataset?

Forecast: 57.1 %

71 forecasters



Jacob Steinhardt, IPCC

It’s not all about the (base) model

Platforms

Overview Documentation API reference

Systems

ChatGPT PLUS

DALL·E

Models

Whisper

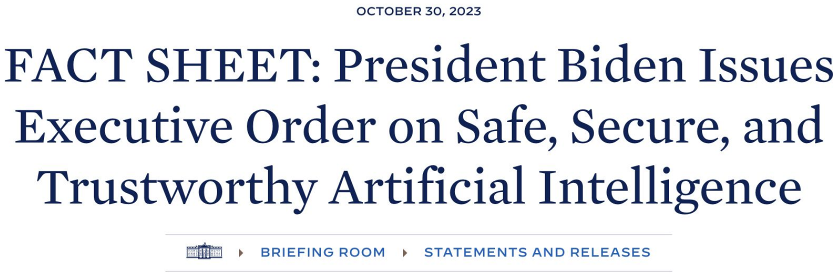
GPT-4

Use cases

casetext

duolingo

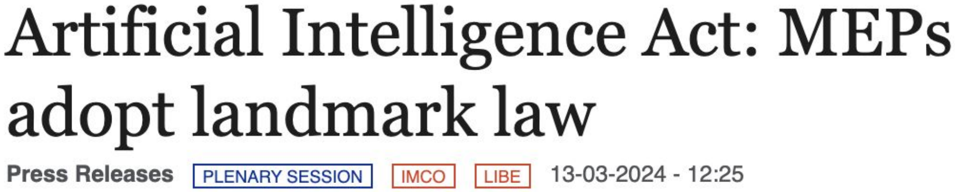
We're transitioning from self-regulation to "real" regulation



OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM ▸ STATEMENTS AND RELEASES



Artificial Intelligence Act: MEPs adopt landmark law

Press Releases PLENARY SESSION IMCO LIBE 13-03-2024 - 12:25

But there will be gaps for the foreseeable future, where the behavior of people within companies makes a difference

(Will return to this at the end)

Tensions between different objectives

Spoiler: I don't think we're on, or even near the Pareto frontier for all of these!

Often the tradeoffs are overstated, and a great thing about AI is that it can help "grow the pie" through automated labor.

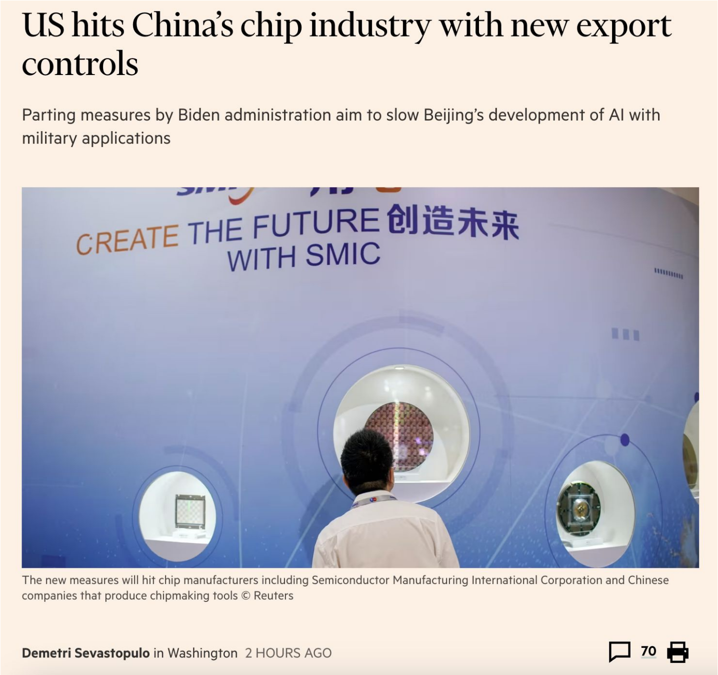
The point is just that there are sometimes actually tradeoffs.

Tension: preventing risks vs. unlocking benefits

Where the rubber meets the road:

- The general pace of progress
- Use case policies / fine-tuning approaches
- General permissiveness of regulation
- Open source

Tension: competing with other companies/countries economically and militarily while collaborating on shared safety challenges

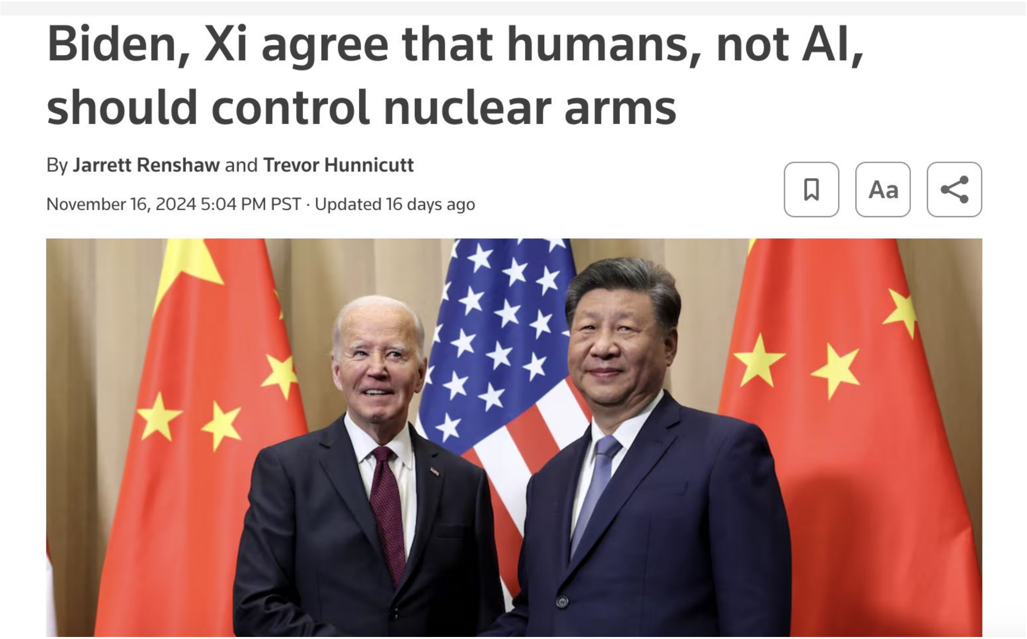


US hits China's chip industry with new export controls

Parting measures by Biden administration aim to slow Beijing's development of AI with military applications

The new measures will hit chip manufacturers including Semiconductor Manufacturing International Corporation and Chinese companies that produce chipmaking tools © Reuters

Demetri Sevastopulo in Washington 2 HOURS AGO



Biden, Xi agree that humans, not AI, should control nuclear arms

By **Jarrett Renshaw** and **Trevor Hunnicutt**

November 16, 2024 5:04 PM PST · Updated 16 days ago

Tension: preventing risks vs. concentrating power

- Open source has a lot of potential for misuse…
    - …but also decentralizes “deep” access to AI capabilities
- Fine-tuning of closed models can reduce direct misuse risk…
    - …but also represents an imposition of a certain set of values

Tension: addressing existing vs. anticipated issues

- There are synergies, but also tradeoffs in policymaker attention, compute, researcher/engineer time, etc.
- E.g. bias/hallucination/"mundane" misuse vs. catastrophic misuse and accidents

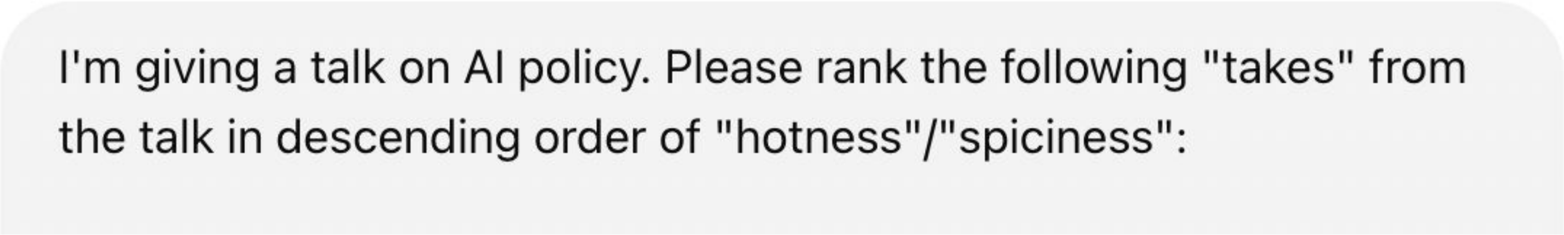
03

Some ~~Hot Spicy Personal~~ Hot Takes

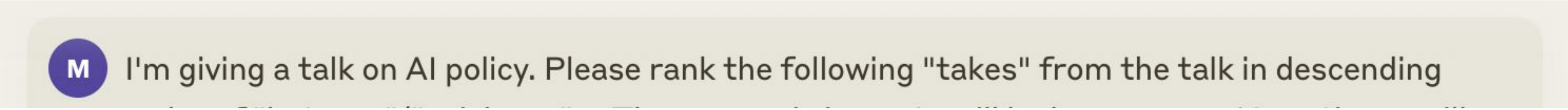
03

Some Hot Takes

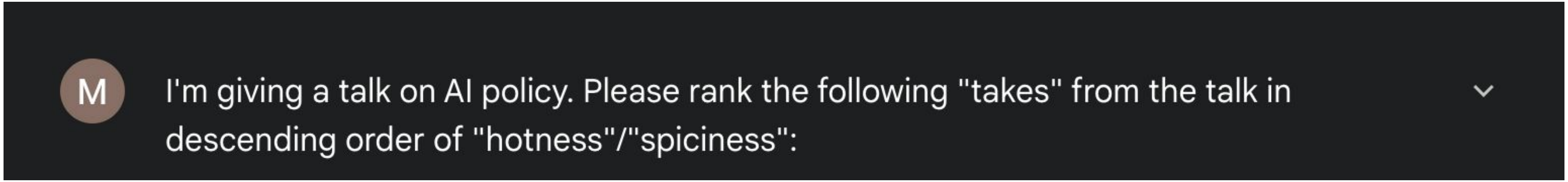
Ranking methodology



I'm giving a talk on AI policy. Please rank the following "takes" from the talk in descending order of "hotness"/"spiciness":



I'm giving a talk on AI policy. Please rank the following "takes" from the talk in descending



I'm giving a talk on AI policy. Please rank the following "takes" from the talk in descending order of "hotness"/"spiciness":

Ranking methodology

Saving one of them, which may or may not be the spiciest, for the section on what you can do

You should watch/read more sci-fi

Battlestar Galactica

Person of Interest

Pantheon

Travelers

Transcendence*

WALL-E

Terminator 2

Westworld

Interstellar

The Diamond Age

Robopocalypse

The Player of Games

Altered Carbon

You should watch/read more sci-fi

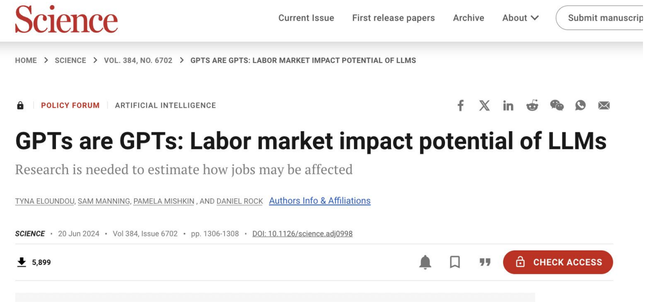
…but not too much, for the reason I’ll give at the end…

The economic impacts will be huge soon

Already, some gig workers are being negatively affected (e.g. copywriters, illustrators)

Others are gaining a lot of productivity, with unclear job consequences

Soon more interactive roles will be impacted, especially ones that are already outsourced/done remotely


Science
Current Issue
First release papers
Archive
About
Submit manuscrip
HOME > SCIENCE > VOL. 384, NO. 6702 > GPTS ARE GPTS: LABOR MARKET IMPACT POTENTIAL OF LLMS
POLICY FORUM | ARTIFICIAL INTELLIGENCE
5,899
CHECK ACCESS

GPTs are GPTs: Labor market impact potential of LLMs

Research is needed to estimate how jobs may be affected

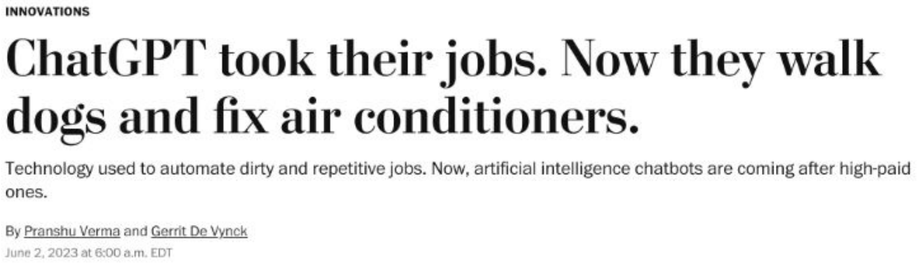
TYNA ELOUNDOU, SAM MANNING, PAMELA MISHKIN , AND DANIEL ROCK Authors Info & Affiliations

SCIENCE • 20 Jun 2024 • Vol 384, Issue 6702 • pp. 1306-1308 • DOI: 10.1126/science.adj0998

Huge != no one has jobs, but…

Double digit annual economic growth within five years

Hundreds of millions displaced from their previous jobs



INNOVATIONS

ChatGPT took their jobs. Now they walk dogs and fix air conditioners.

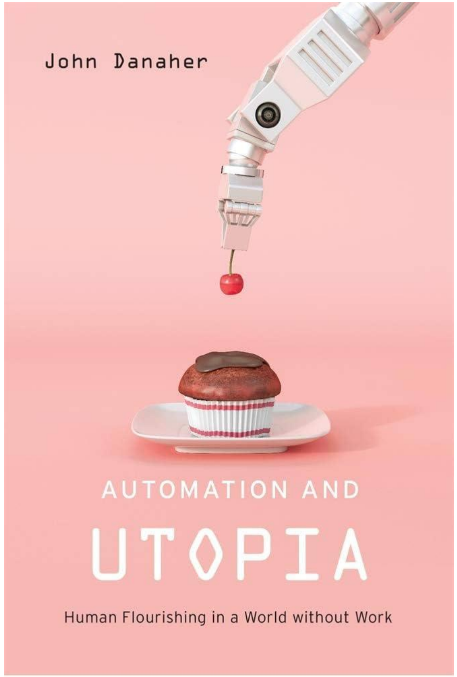
Technology used to automate dirty and repetitive jobs. Now, artificial intelligence chatbots are coming after high-paid ones.

By Pranshu Verma and Gerrit De Vynck

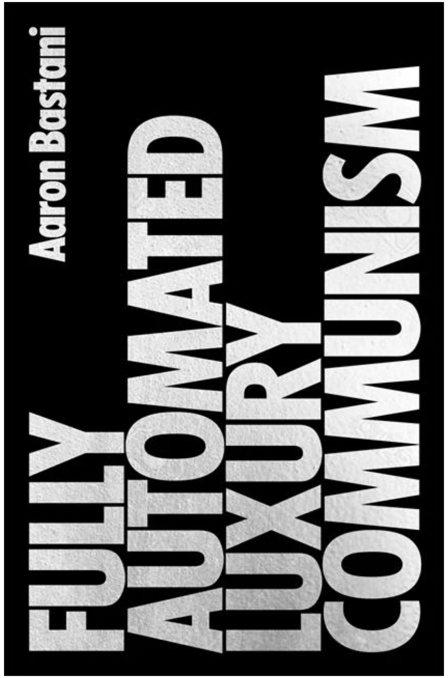
June 2, 2023 at 6:00 a.m. EDT

Washington Post

We need to talk about post-work futures now



free time
julie l. rose



Investing in society's resilience to AI's impacts is ~a free lunch

Not literally free, but –

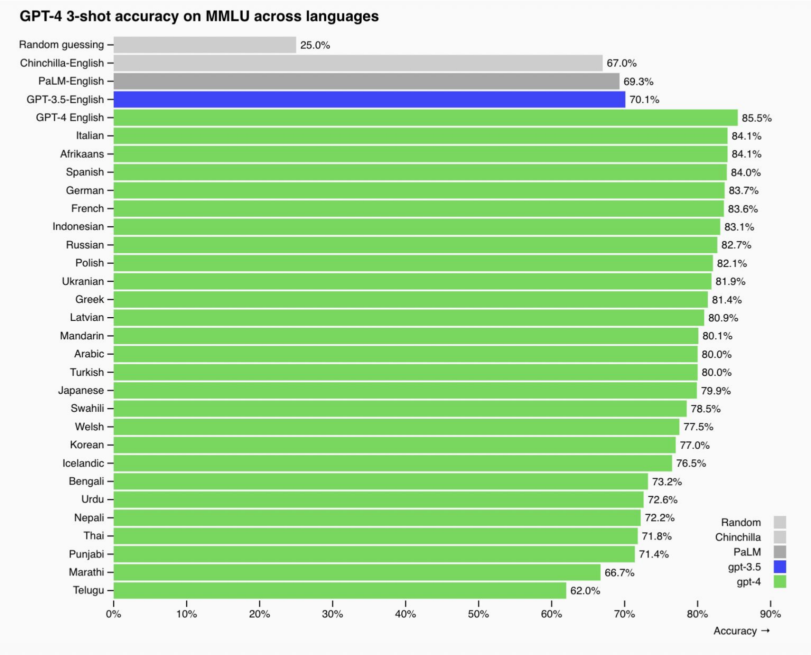
If all countries in the world had strong social safety nets (to cushion people from harms associated with unemployment), robust investment in cybersecurity, ubiquitous physical defenses against biological threats (e.g. far-UVC)...

We could distribute AI capabilities more widely/evenly and have more transparency – right now, these are sometimes in tension with safety/security

Access to AI is unusually equal by the standards of previous technologies

- Piggybacking on the internet → fast distribution
  - Though the digital divide is still a huge issue
- For part of this year, the best free model and the best paid model were the same model (other than rate limits)
- The point here is not to excuse cases where companies fall short! Just to calibrate.

Access to AI is unusually equal by the standards of previous technologies



...but in the worst case, it could concentrate power as never before...

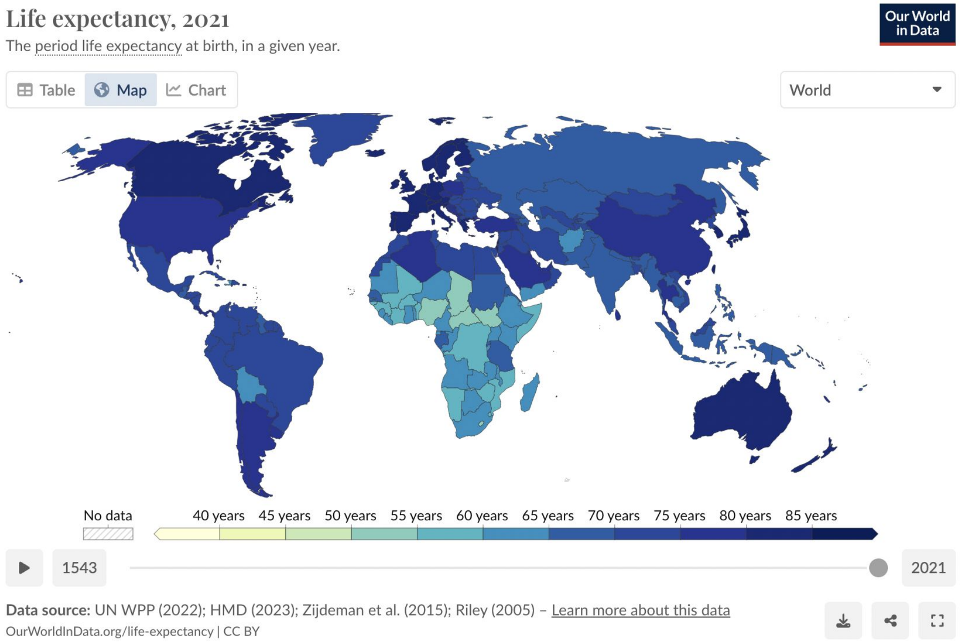
- More compute → better answers → by default, you should expect AI services to bifurcate dramatically between free and paid
- AI can also automate surveillance, censorship, political messaging, etc. at scale with precision

...but it could still be a massive force for equality

...because the world of today is a very weak baseline on this score...

...and because there may be diminishing returns on cognition in many aspects of life.

...but it could still be a massive force for equality



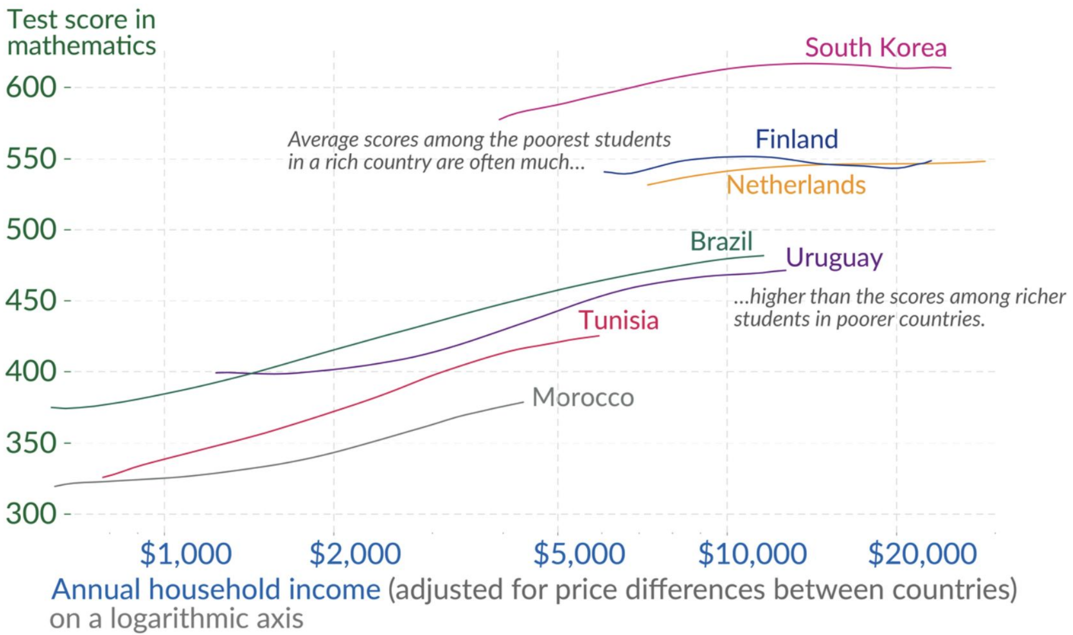
Life expectancy, 2021

The period life expectancy at birth, in a given year.

Data source: UN WPP (2022); HMD (2023); Zijdeman et al. (2015); Riley (2005) – Learn more about this data

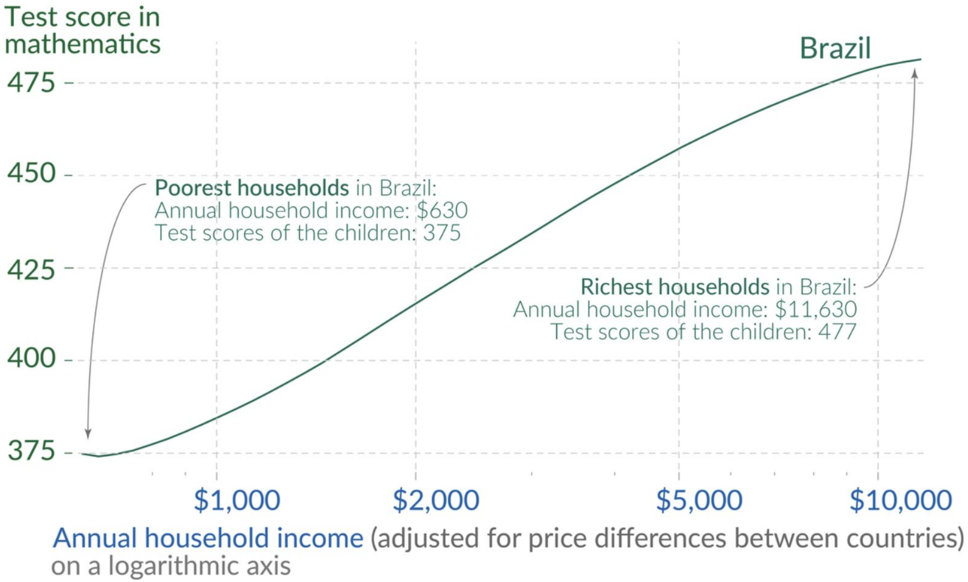
OurWorldInData.org/life-expectancy | CC BY

...but it could still be a massive force for equality



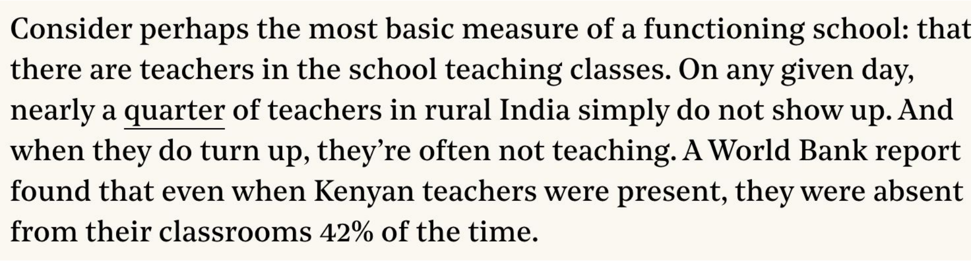
Our World in Data

...but it could still be a massive force for equality



Our World in Data

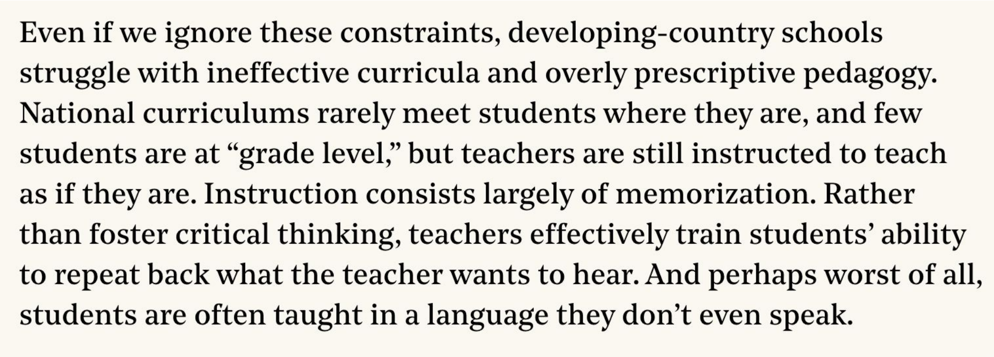
...but it could still be a massive force for equality



Consider perhaps the most basic measure of a functioning school: that there are teachers in the school teaching classes. On any given day, nearly a quarter of teachers in rural India simply do not show up. And when they do turn up, they're often not teaching. A World Bank report found that even when Kenyan teachers were present, they were absent from their classrooms 42% of the time.

Low quantity of education

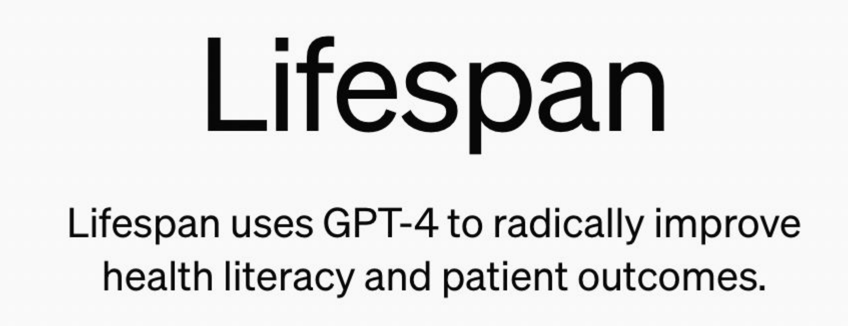
Low quality of education



Even if we ignore these constraints, developing-country schools struggle with ineffective curricula and overly prescriptive pedagogy. National curriculums rarely meet students where they are, and few students are at "grade level," but teachers are still instructed to teach as if they are. Instruction consists largely of memorization. Rather than foster critical thinking, teachers effectively train students' ability to repeat back what the teacher wants to hear. And perhaps worst of all, students are often taught in a language they don't even speak.

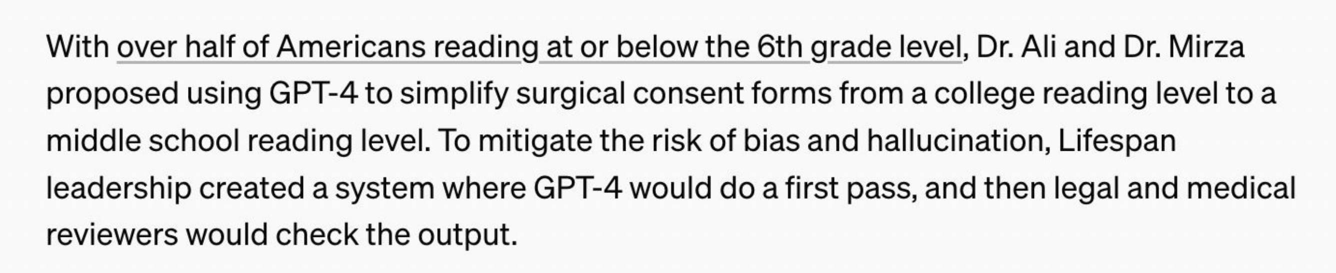
Gilbert, 2024

...but it could still be a massive force for equality



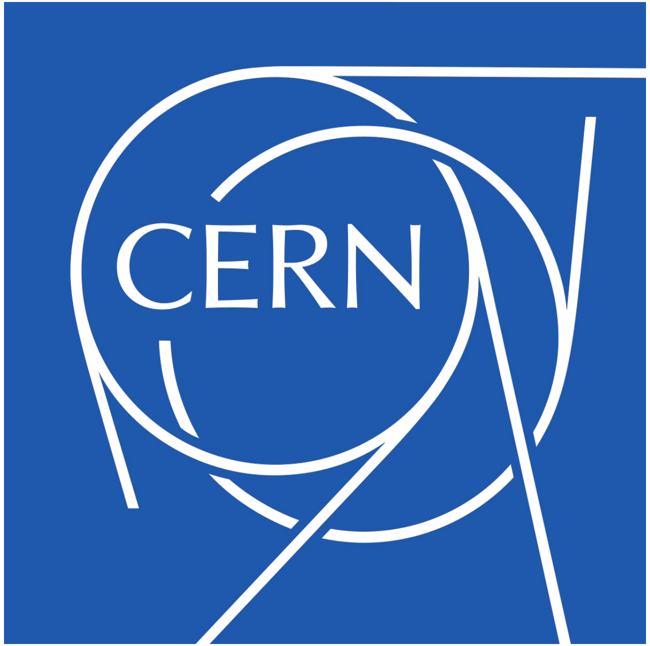
Lifespan

Lifespan uses GPT-4 to radically improve health literacy and patient outcomes.



With over half of Americans reading at or below the 6th grade level, Dr. Ali and Dr. Mirza proposed using GPT-4 to simplify surgical consent forms from a college reading level to a middle school reading level. To mitigate the risk of bias and hallucination, Lifespan leadership created a system where GPT-4 would do a first pass, and then legal and medical reviewers would check the output.

There should be much more serious consideration of a “CERN for AI” scenario



“The CERN approach” ~ = pooling resources to build and operate centralized infrastructure in a transparent way, as a global scientific community, for civilian rather than military purposes

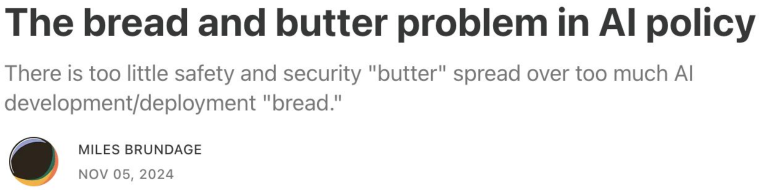
There should be much more serious consideration of a “CERN for AI” scenario

x 20 vs. 1 x

There should be much more serious consideration of a “CERN for AI” scenario

I feel thin, sort of stretched, like butter scraped over too much bread.

Lord of the Rings: Fellowship of the Ring



The bread and butter problem in AI policy

There is too little safety and security "butter" spread over too much AI development/deployment "bread."

MILES BRUNDAGE

NOV 05, 2024

There should be much more serious consideration of a “CERN for AI” scenario

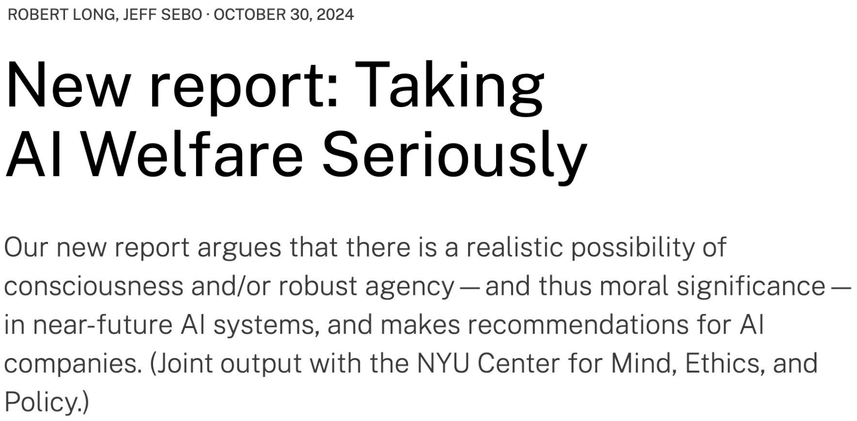
- Ensure that the very most capable models are developed extremely securely and safely:
  - Combine the world’s talent on security, then safety, then capabilities – in that order, otherwise you just speed up development and everyone steals it and fine-tunes it in dangerous ways
- When models are derisked, distribute and deploy them widely

There should be much more serious consideration of a “CERN for AI” scenario

- It’s not obvious that this is the right thing to do but it deserves serious debate and being fleshed out
- Key question: how could this be designed such that there is *distributed* control over these *centralized* capabilities (e.g. multiple parties can stop a dangerous training/inference run).
  - This is partly a technical and partly a political question.

AI sentience will also be a huge issue

The costs of error in either direction are huge



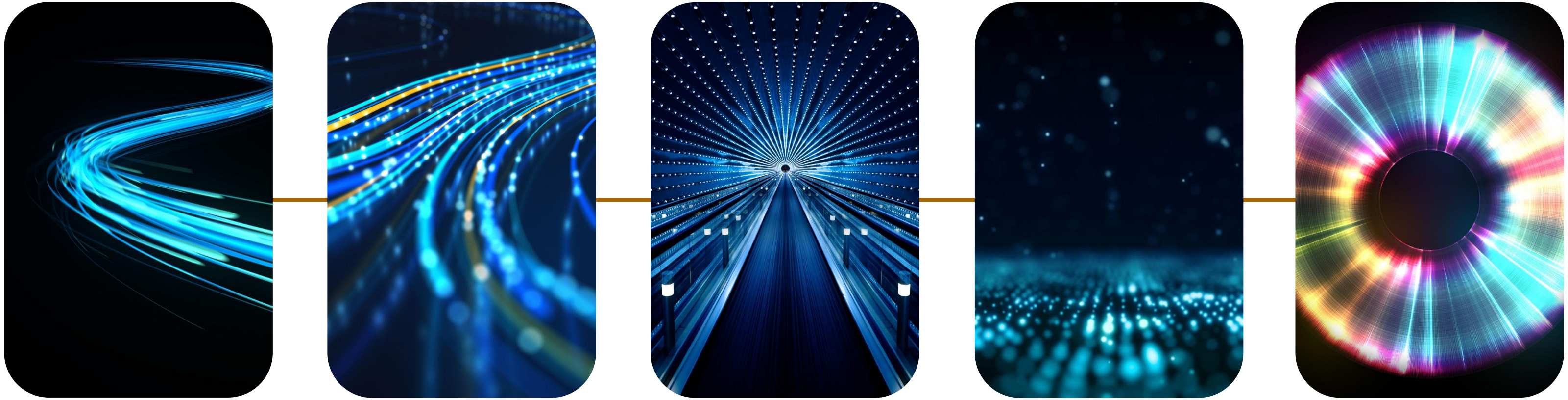
ROBERT LONG, JEFF SEBO · OCTOBER 30, 2024

New report: Taking AI Welfare Seriously

Our new report argues that there is a realistic possibility of consciousness and/or robust agency — and thus moral significance — in near-future AI systems, and makes recommendations for AI companies. (Joint output with the NYU Center for Mind, Ethics, and Policy.)

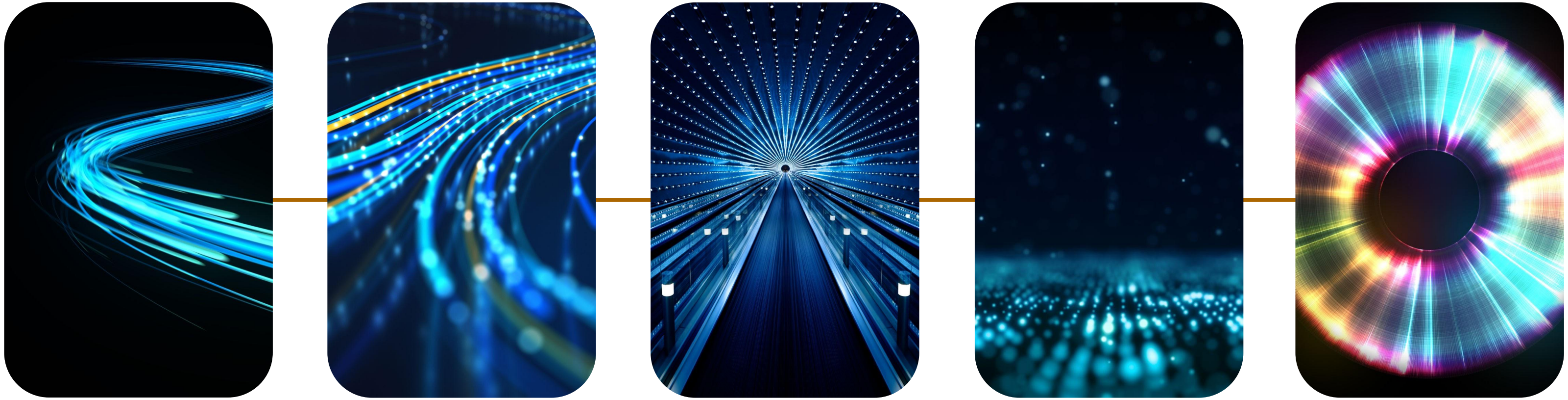
04

Where You Fit In



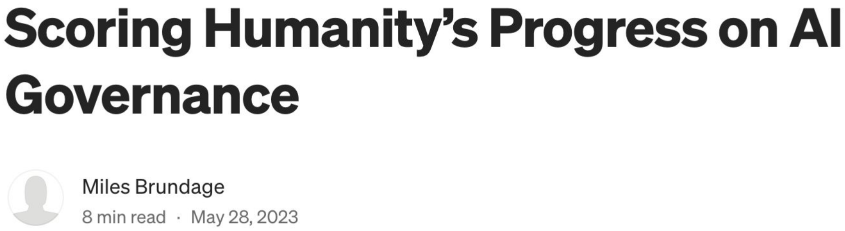
Even if you don’t care about AI policy, AI policy cares about you.

Public (government) and private (corporate/non-profit etc.) decisions will affect your career in various ways, as well as your life as a citizen more generally.



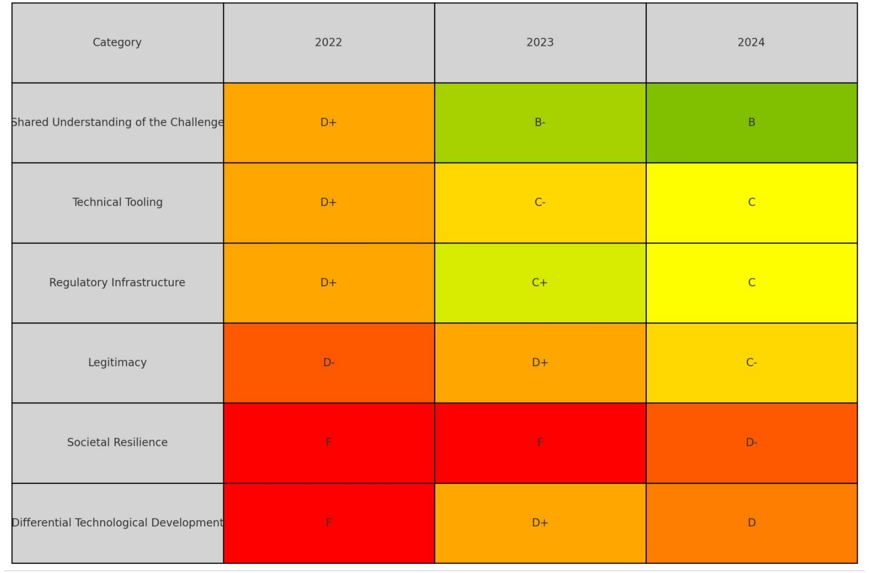
See also my blog post: “FAQs and General Advice on AI Policy Careers”

We're running out of time



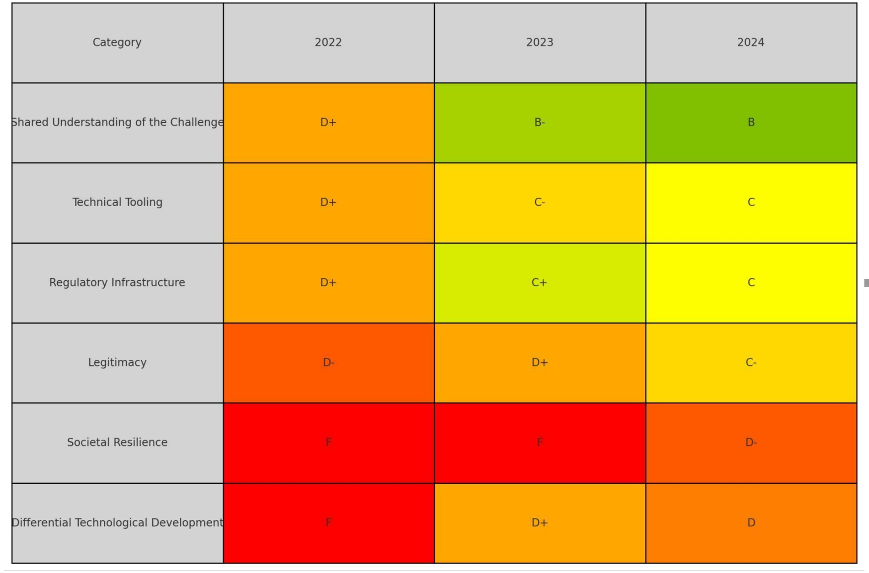
Scoring Humanity's Progress on AI Governance

Miles Brundage
8 min read · May 28, 2023



Category	2022	2023	2024
Shared Understanding of the Challenge	D+	B-	B
Technical Tooling	D+	C-	C
Regulatory Infrastructure	D+	C+	C
Legitimacy	D-	D+	C-
Societal Resilience	F	F	D-
Differential Technological Development	F	D+	D

We’re running out of time

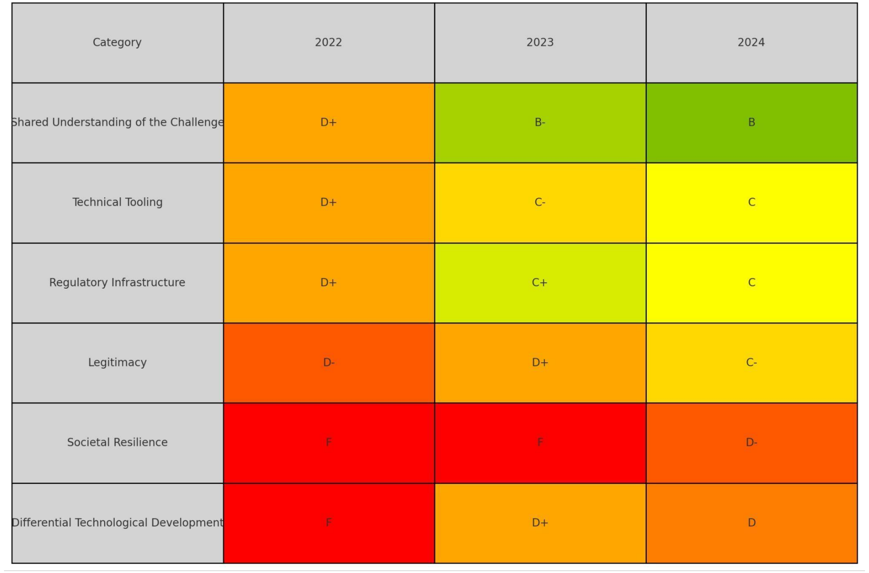


Category	2022	2023	2024
Shared Understanding of the Challenge	D+	B-	B
Technical Tooling	D+	C-	C
Regulatory Infrastructure	D+	C+	C
Legitimacy	D-	D+	C-
Societal Resilience	F	F	D-
Differential Technological Development	F	D+	D

2025 2026 2027...

?

We’re running out of time



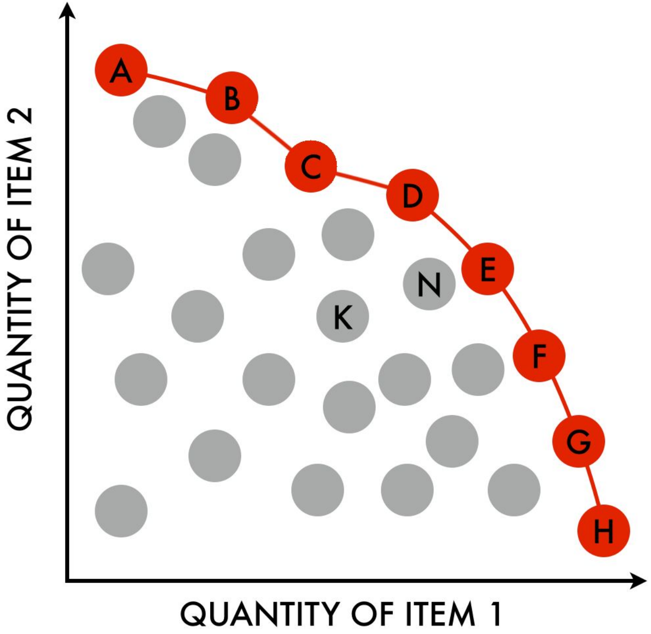
Category	2022	2023	2024
Shared Understanding of the Challenge	D+	B-	B
Technical Tooling	D+	C-	C
Regulatory Infrastructure	D+	C+	C
Legitimacy	D-	D+	C-
Societal Resilience	F	F	D-
Differential Technological Development	F	D+	D

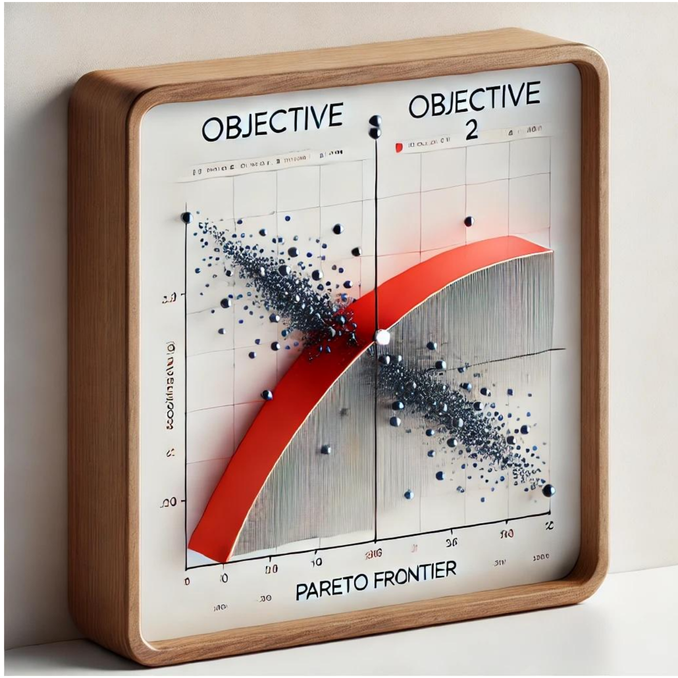
Your career may not (have to) be as long as you thought

Not sure now is the best time to start a PhD...

...except, perhaps, if you’re OK multitasking

The AI policy Pareto frontier





Wikipedia/DALL-E 3

The AI policy Pareto frontier

Maximizing benefits and minimizing risks…

…privacy/copyright protection vs. knowledge about the world…

…raising the ceiling of capabilities/raising the floor of access

Ways you can help

Pushing out the frontier with technical innovation

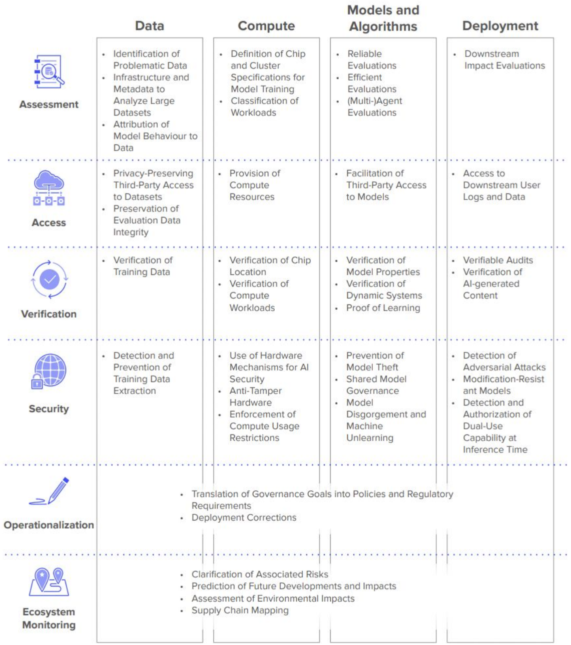
Helping organizations and the world make informed decisions about how to get to the frontier and pick a point on it

Pushing out the frontier with technical innovation

Open Problems in Technical AI Governance

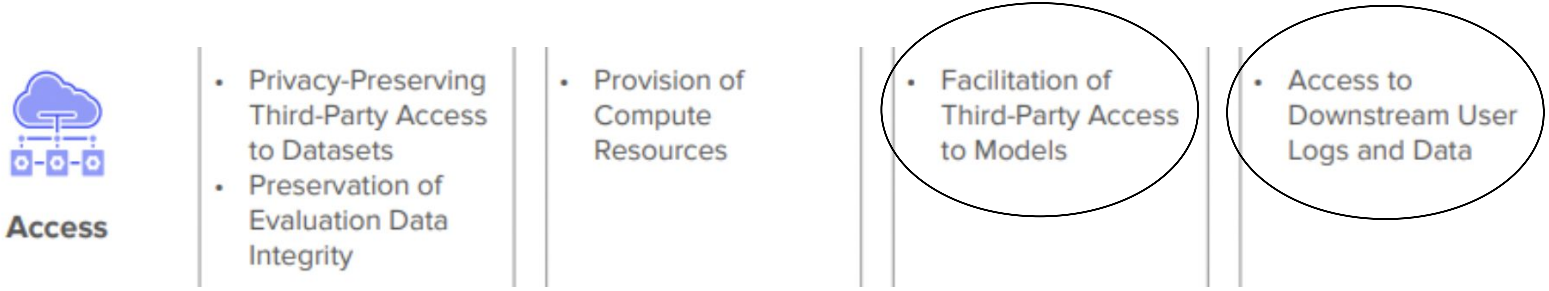
Anka Reuel* *Stanford University* *anka.reuel@stanford.edu*
Ben Bucknall* *Centre for the Governance of AI &*
Oxford Martin AI Governance Initiative *ben.bucknall@governance.ai*

Stephen Casper *MIT CSAIL*
Tim Fist *Institute for Progress & Center for a New American Security*
Lisa Soder *interface – Tech Analysis and Policy Ideas for Europe e.V.*
Onni Aarne *Institute for AI Policy and Strategy*
Lewis Hammond *University of Oxford & Cooperative AI Foundation*
Lujain Ibrahim *University of Oxford*
Alan Chan *Centre for the Governance of AI & Mila*
Peter Wills *Centre for the Governance of AI & University of Oxford*
Markus Anderljung *Centre for the Governance of AI*
Ben Garfinkel *Centre for the Governance of AI*
Lennart Heim *Centre for the Governance of AI*
Andrew Trask *OpenMined & University of Oxford*
Gabriel Mukobi *Stanford University*
Rylan Schaeffer *Stanford University*
Mauricio Baker *Independent Researcher*
Sara Hooker *Cohere For AI*
Irene Solaiman *Hugging Face*
Alexandra Sasha Luccioni *Hugging Face*
Nitarshan Rajkumar *University of Cambridge*
Nicolas Moës *The Future Society*
Neel Guha *Stanford University*
Jessica Newman *University of California, Berkeley*
Yoshua Bengio *University of Montreal & Mila*
Tobin South *MIT*
Alex Pentland *Stanford HAI*
Jeffrey Ladish *Palisade Research*
Sanmi Koyejo *Stanford University, Virtue AI*
Mykel J. Kochenderfer *Stanford University*
Robert Trager *Oxford Martin AI Governance Initiative, Blavatnik School of Government &*
University of Oxford



	Data	Compute	Models and Algorithms	Deployment
Assessment	• Identification of Problematic Data • Infrastructure and Metadata to Analyze Large Datasets • Attribution of Model Behaviour to Data	• Definition of Chip and Cluster Specifications for Model Training • Classification of Workloads	• Reliable Evaluations • Efficient Evaluations • (Multi-)Agent Evaluations	• Downstream Impact Evaluations
Access	• Privacy-Preserving Third-Party Access to Datasets • Preservation of Evaluation Data Integrity	• Provision of Compute Resources	• Facilitation of Third-Party Access to Models	• Access to Downstream User Logs and Data
Verification	• Verification of Training Data	• Verification of Chip Location • Verification of Compute Workloads	• Verification of Model Properties • Verification of Dynamic Systems • Proof of Learning	• Verifiable Audits • Verification of AI-generated Content
Security	• Detection and Prevention of Training Data Extraction	• Use of Hardware Mechanisms for AI Security • Anti-Tamper Hardware • Enforcement of Compute Usage Restrictions	• Prevention of Model Theft • Shared Model Governance • Model Disgorgement and Machine Unlearning	• Detection of Adversarial Attacks • Modification-Resistant Models • Detection and Authorization of Dual-Use Capability at Inference Time
Operationalization	• Translation of Governance Goals into Policies and Regulatory Requirements • Deployment Corrections			
Ecosystem Monitoring	• Clarification of Associated Risks • Prediction of Future Developments and Impacts • Assessment of Environmental Impacts • Supply Chain Mapping			

Pushing out the frontier with technical innovation



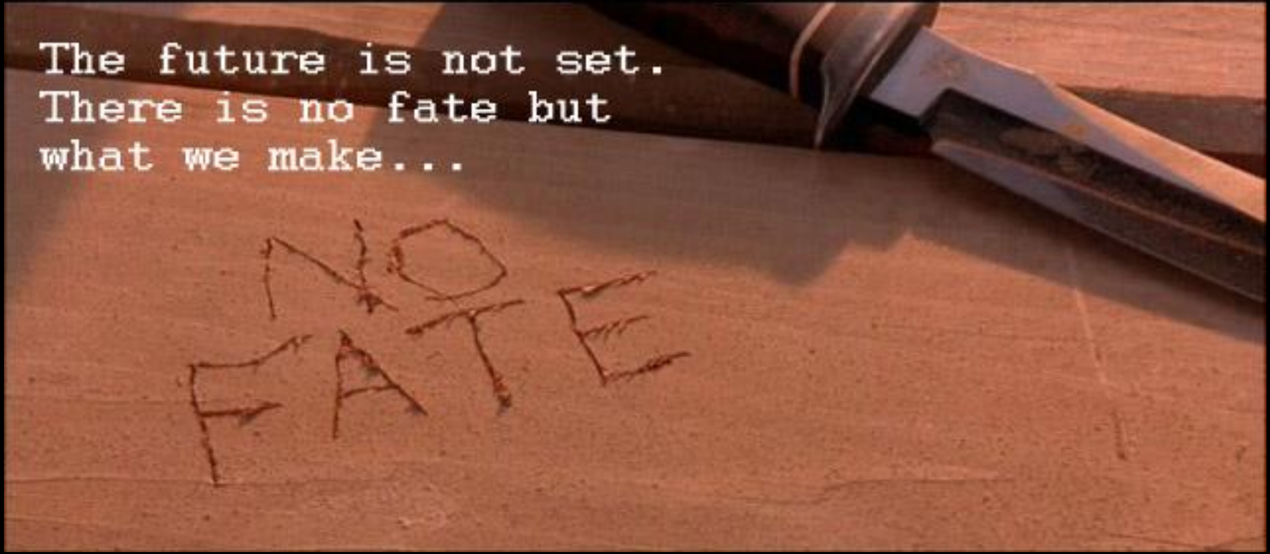
Access	- Privacy-Preserving Third-Party Access to Datasets - Preservation of Evaluation Data Integrity	- Provision of Compute Resources	- Facilitation of Third-Party Access to Models	- Access to Downstream User Logs and Data

Getting to the frontier and moving around on it: your voice matters!

- Companies have, and may always have, some latitude in how they interpret voluntary commitments and regulations.
    - This is good in some respects by fostering innovation, but also can enable recklessness
- Companies will generally want you to be happy, which gives you power.
    - You won't always get your way (company behavior isn't *just* a function of employee views – there are also competitive pressures, etc.), but every bit of thoughtfulness contributes to company culture

Getting to the frontier and moving around on it: your voice matters!

- Public and private debates matter: it can be helpful to push back when someone is mischaracterizing their opponents, dismissing legitimate concerns, etc.
  - Though don’t spend all your time on it - there’s a lot of real work to do, as well



Terminator 2: Judgment Day

Acknowledgments

Policy is a team sport. My views have been shaped by everyone I’ve worked with, at OpenAI and beyond, as well as my friends, family, and partner. Errors and views are my own.

Thank you!

@miles_brundage on Twitter
Miles Brundage on Google Scholar
milesbrundage.substack.com