CORNELL TECH

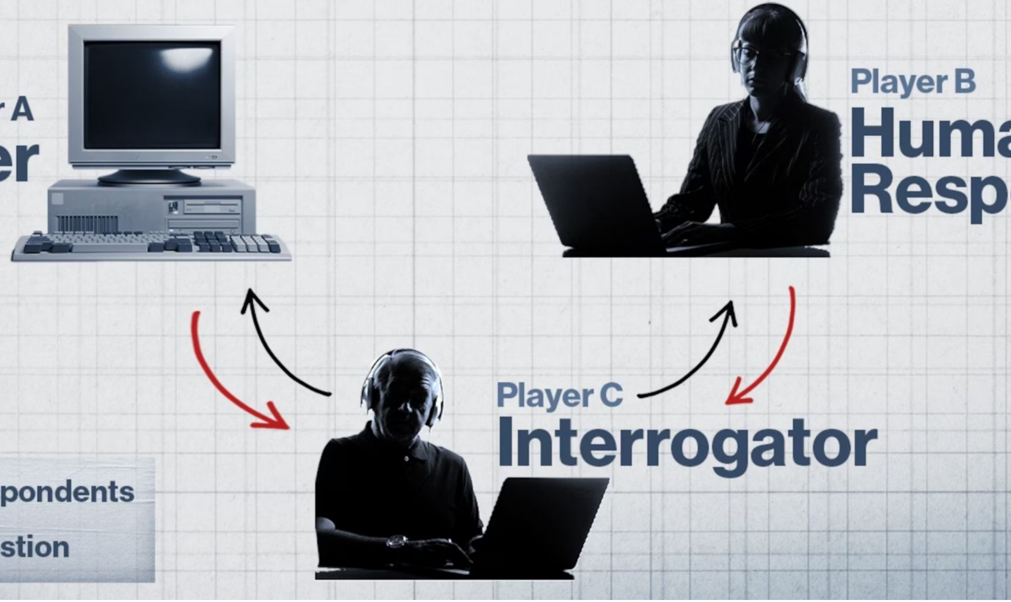
CS 5434 | Fall 2025 | Trustworthy AI

AI Governance & AGI

12/8/2025

Turing Test:

During the Turing Test, the human interrogator asks several questions to both players. Based on the answers, the interrogator attempts to determine which player is a computer and which player is a human respondent.



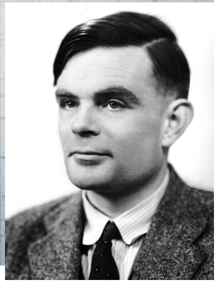
Player A
Computer

Player B
Human Responder

Player C
Interrogator

■ Question to Respondents

🟥 Answers to Question



Why AGI?

1. The brain is a computational machine.

2. We can simulate any computational machine using a computer.

3. We can (eventually) build a brain.

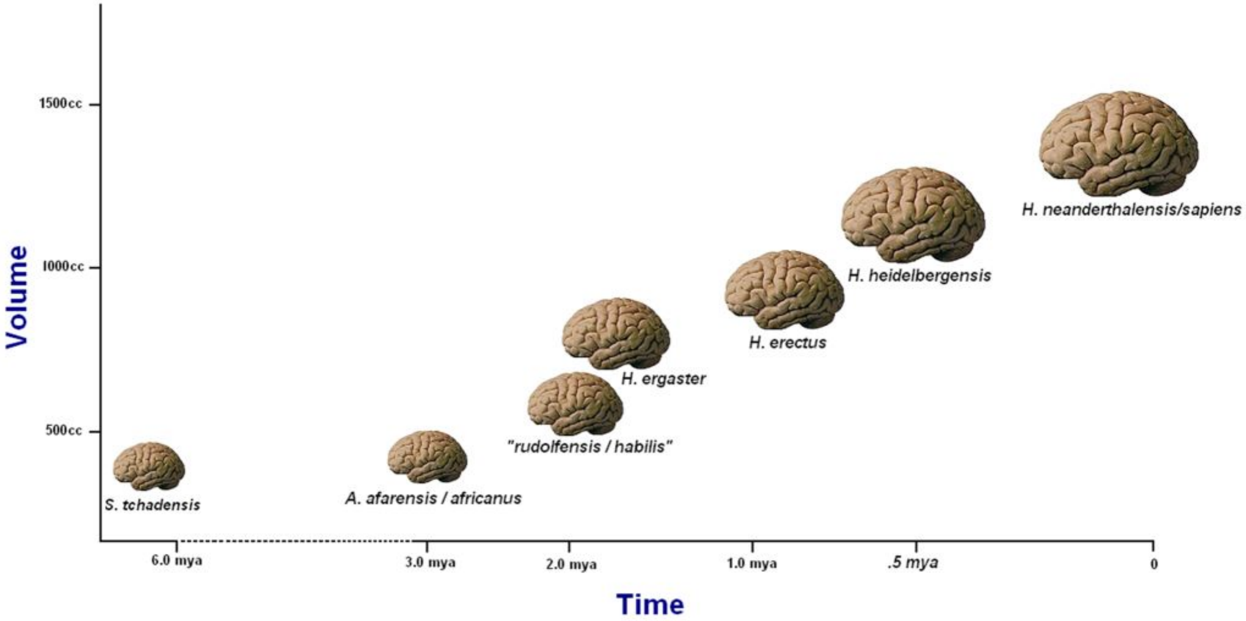
Daniel Dennett

Moravec’s paradox

"it is comparatively easy to make computers exhibit adult level performance on intelligence tests or playing checkers, and difficult or impossible to give them the skills of a one-year-old when it comes to perception and mobility" – Moravec (1988)

This is why robots learned to play chess before they could walk

Big takeaway: AI capability frontier is <u>counterintuitive</u> and <u>difficult to predict</u>

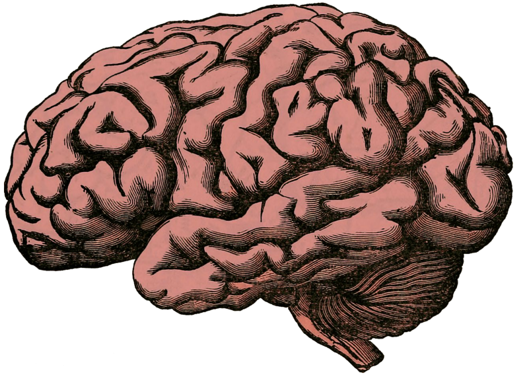


How many FLOPs does a human brain use?

LED lightbulb

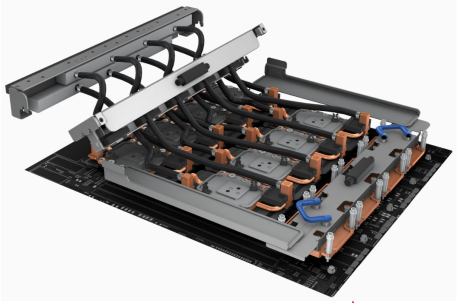
12 watts

Human brain



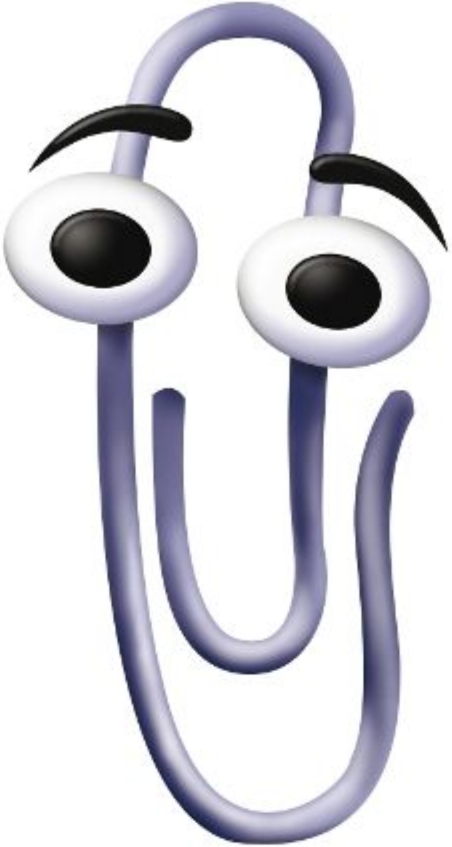
20 watts

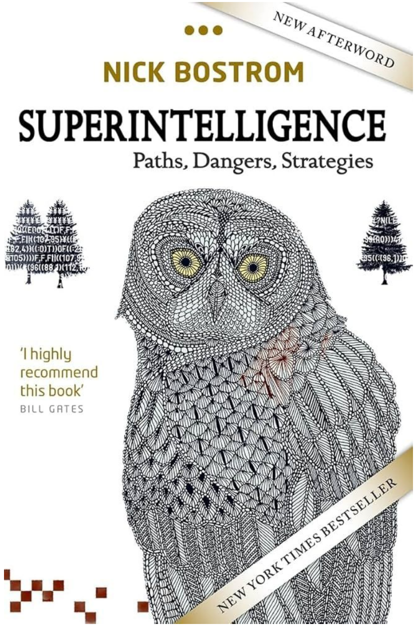
NVIDIA B200 GPU



1000 watts

GPT-5 pretraining used ~50,000 of these for 3-6 months





NEW AFTERWORD

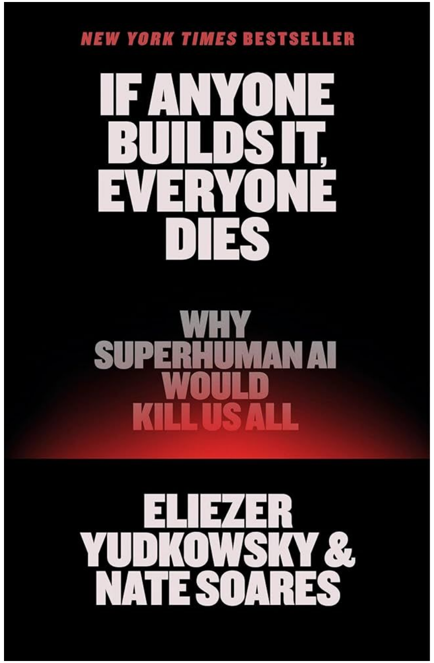
NICK BOSTROM

SUPERINTELLIGENCE

Paths, Dangers, Strategies

'I highly recommend this book'
BILL GATES

NEW YORK TIMES BESTSELLER

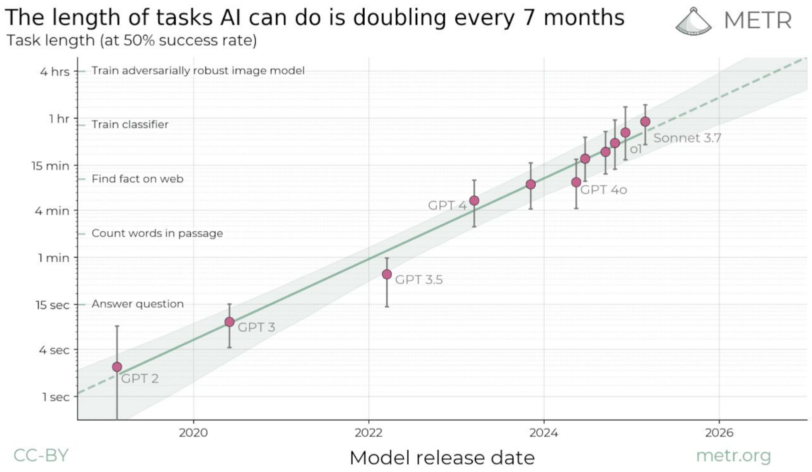


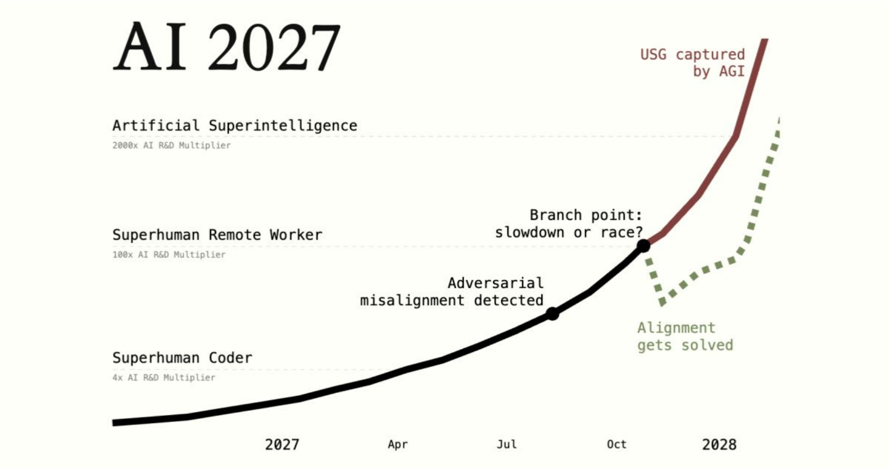
NEW YORK TIMES BESTSELLER

IF ANYONE BUILDS IT, EVERYONE DIES

WHY SUPERHUMAN AI WOULD KILL US ALL

ELIEZER YUDKOWSKY & NATE SOARES

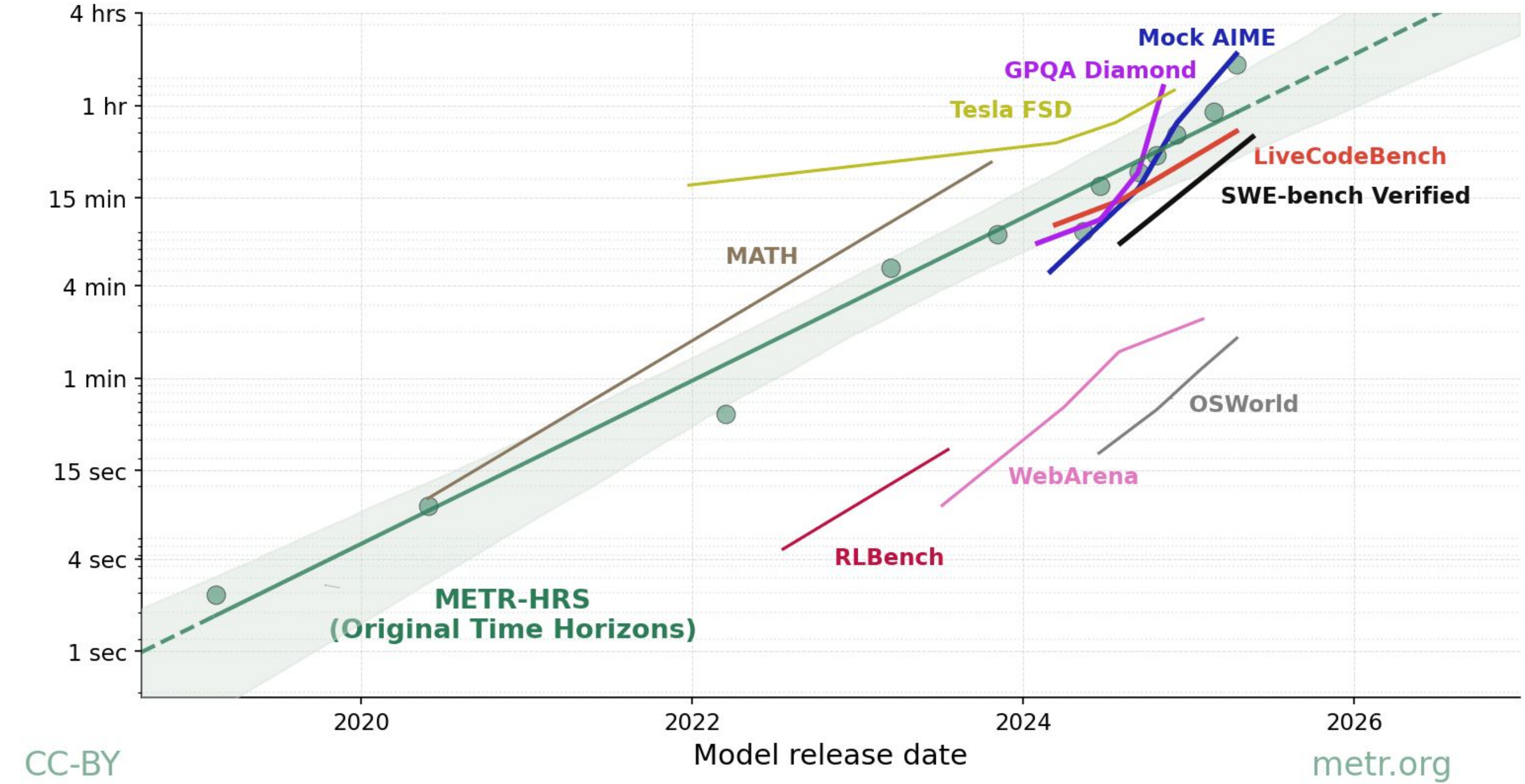

The length of tasks AI can do is doubling every 7 months
METR
Task length (at 50% success rate)
4 hrs
1 hr
15 min
4 min
1 min
15 sec
4 sec
1 sec
Train adversarially robust image model
Train classifier
Find fact on web
Count words in passage
Answer question
GPT 2
GPT 3
GPT 3.5
GPT 4
GPT 4o
o1
Sonnet 3.7
2020
2022
2024
2026
Model release date
CC-BY
metr.org


AI 2027
USG captured
by AGI
Artificial Superintelligence
2000x AI R&D Multiplier
Superhuman Remote Worker
100x AI R&D Multiplier
Branch point:
slowdown or race?
Adversarial
misalignment detected
Alignment
gets solved
Superhuman Coder
4x AI R&D Multiplier
2027
Apr
Jul
Oct
2028

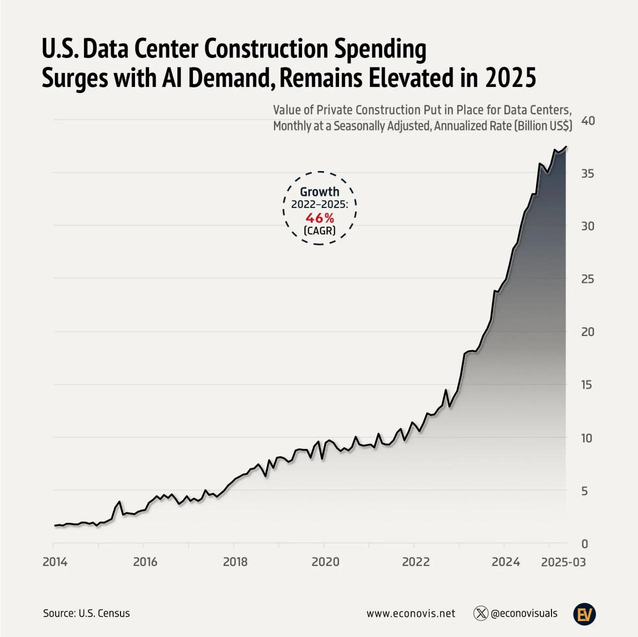
AI time horizons are increasing in many domains

Task length (at 50% success rate)

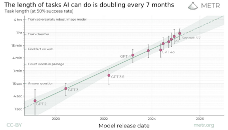
METR



metr.org



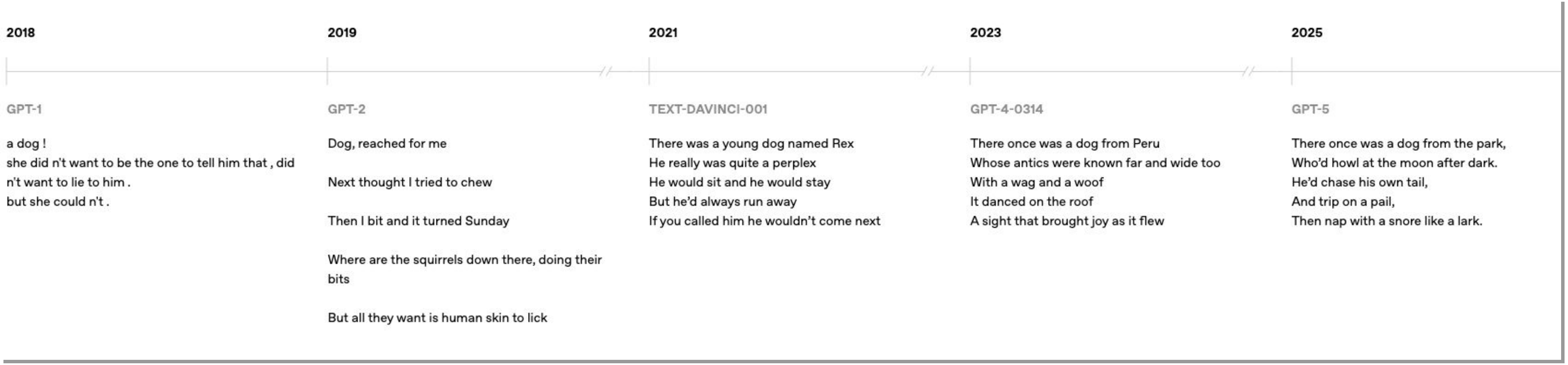
U.S. Data Center Construction Spending
Surges with AI Demand, Remains Elevated in 2025
Value of Private Construction Put in Place for Data Centers, Monthly at a Seasonally Adjusted, Annualized Rate (Billion US$)
Growth 2022–2025: 46% (CAGR)
40
35
30
25
20
15
10
5
0
2014
2016
2018
2020
2022
2024
2025-03
Source: U.S. Census
www.econovis.net
@econovisuals
EV



The length of tasks AI can do is doubling every 7 months
METR
Task length (at 50% success rate)
4 hrs
1 hr
15 min
4 min
1 min
15 sec
4 sec
1 sec
Train adversarially robust image model
Train classifier
Find fact on web
Count words in passage
Answer question
GPT 2
GPT 3
GPT 3.5
GPT 4
GPT 4o
o1
Sonnet 3.7
2020
2022
2024
2026
Model release date
CC-BY
metr.org

Backdrop: AI Capabilities

Write a limerick about a dog:



2018

GPT-1

a dog !
she did n't want to be the one to tell him that , did n't want to lie to him .
but she could n't .

2019

GPT-2

Dog, reached for me

Next thought I tried to chew

Then I bit and it turned Sunday

Where are the squirrels down there, doing their bits

But all they want is human skin to lick

2021

TEXT-DAVINCI-001

There was a young dog named Rex
He really was quite a perplex
He would sit and he would stay
But he'd always run away
If you called him he wouldn't come next

2023

GPT-4-0314

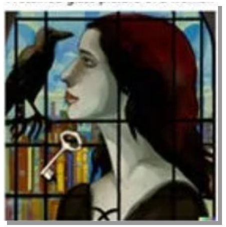
There once was a dog from Peru
Whose antics were known far and wide too
With a wag and a woof
It danced on the roof
A sight that brought joy as it flew

2025

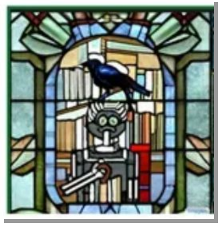
GPT-5

There once was a dog from the park,
Who'd howl at the moon after dark.
He'd chase his own tail,
And trip on a pail,
Then nap with a snore like a lark.

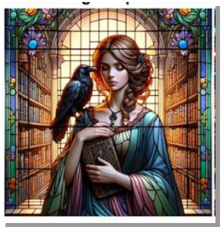
A stained glass picture of a woman in a library with a raven on her shoulder with a key in its mouth



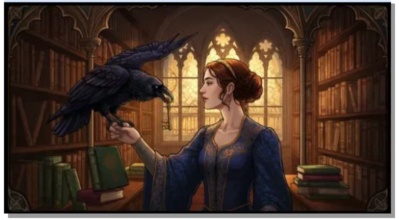
June 2022



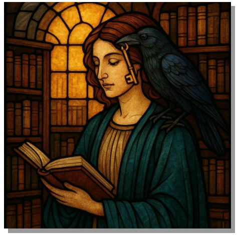
Sep 2022



Jan 2024

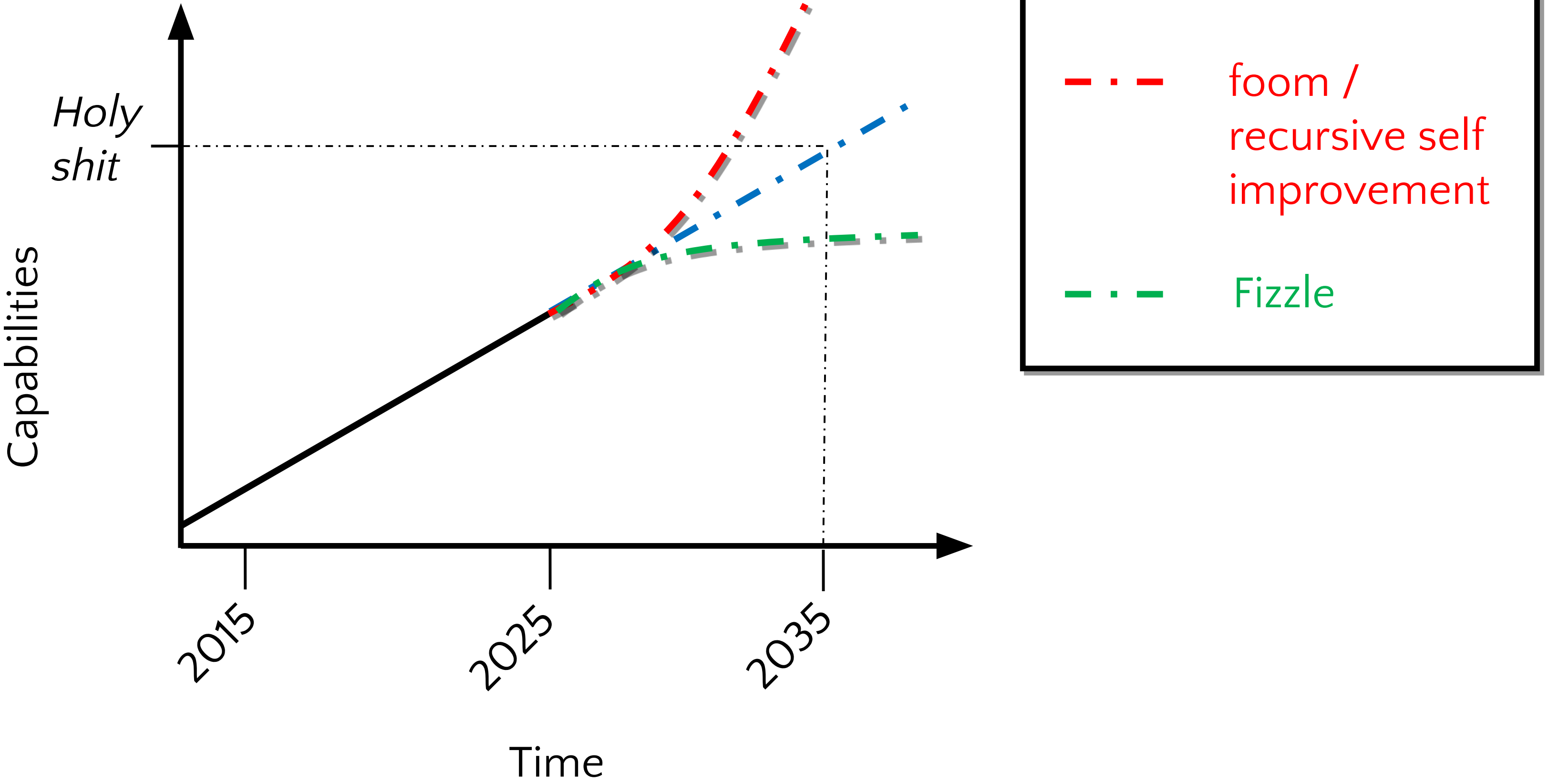


Dec 2024



June 2025

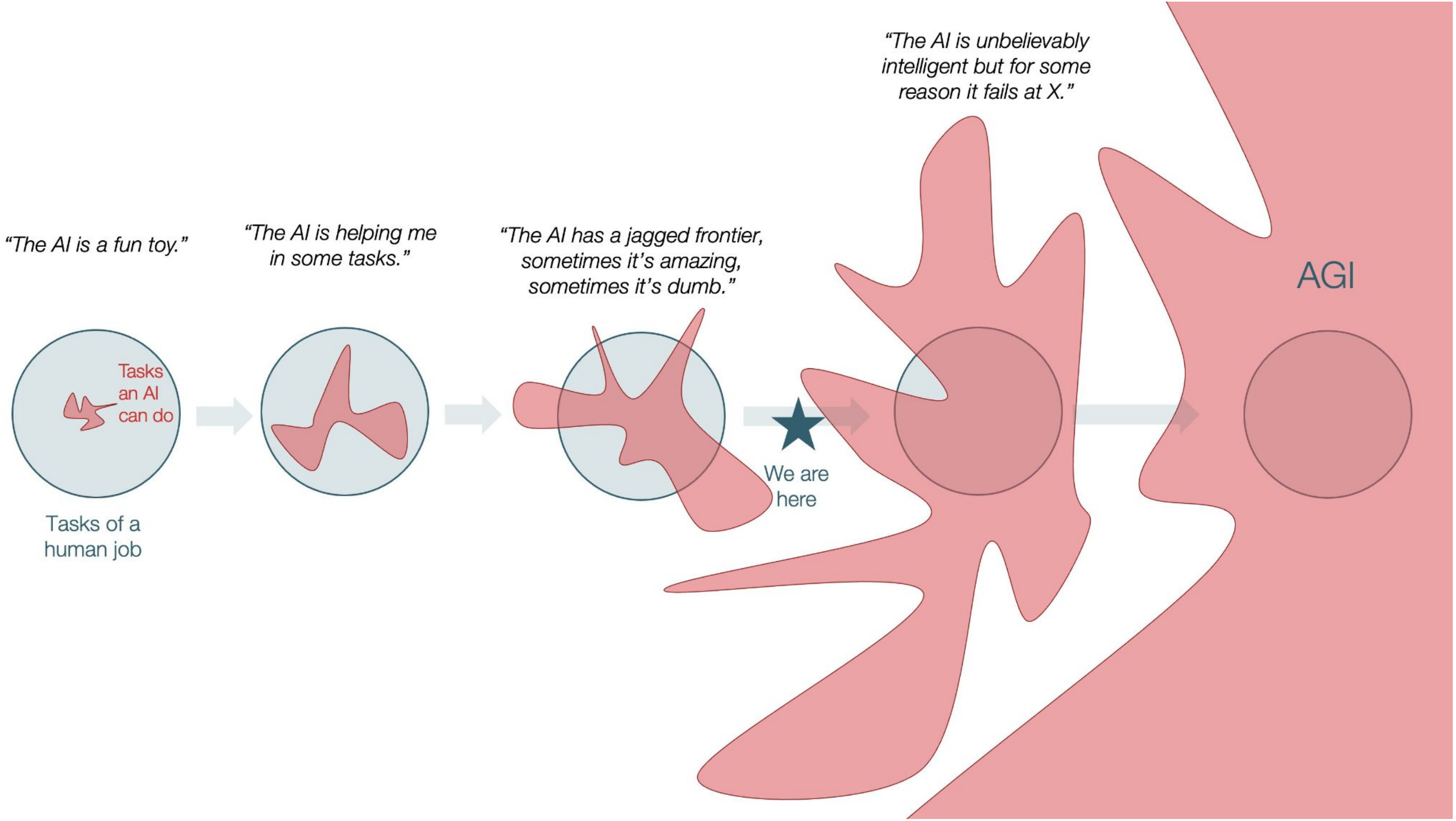
Backdrop: AI capabilities



AI capabilities are “jagged”

- chatGPT can help me optimize the TFLOPS of my matmul CUDA kernel, but still can’t quite plan me a weekend trip

- Newest models got gold medals on the IMO this year, but still make basic errors of arithmetic

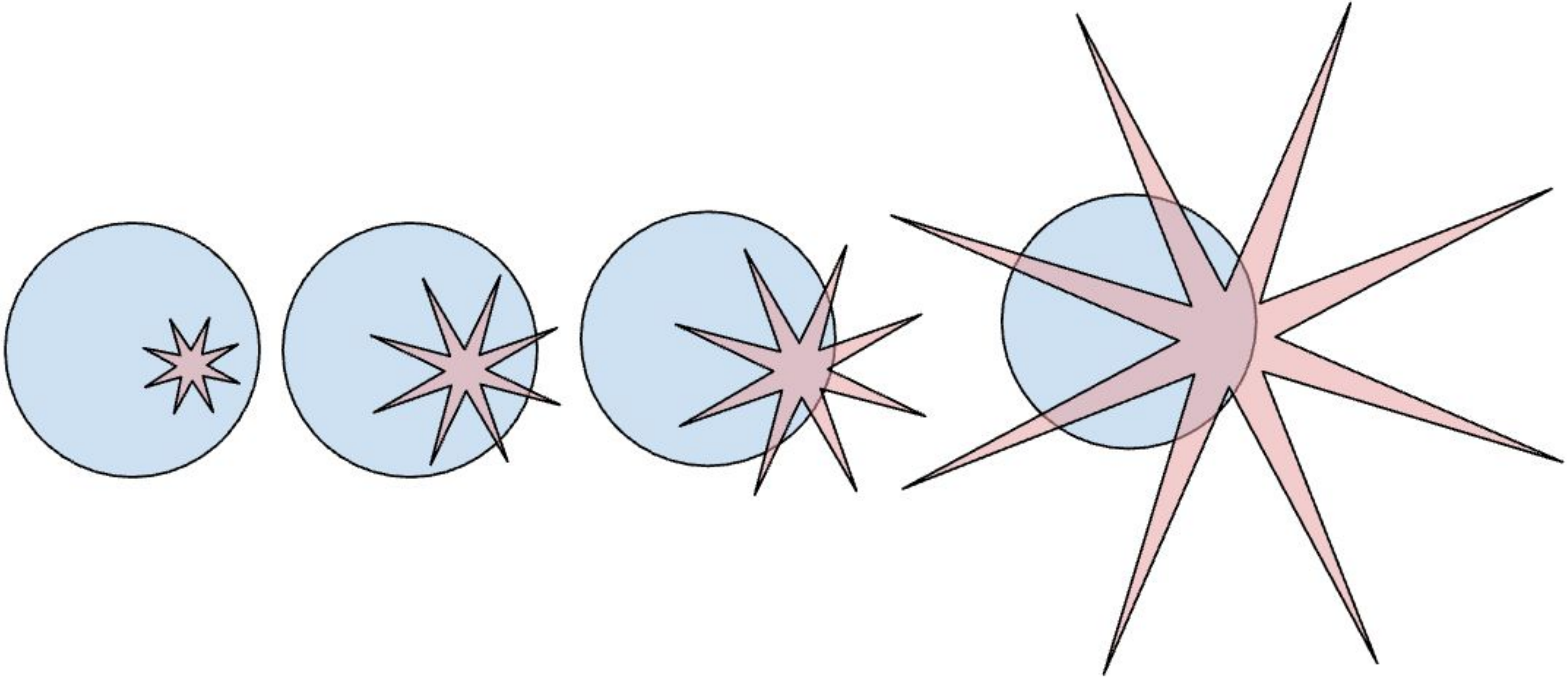


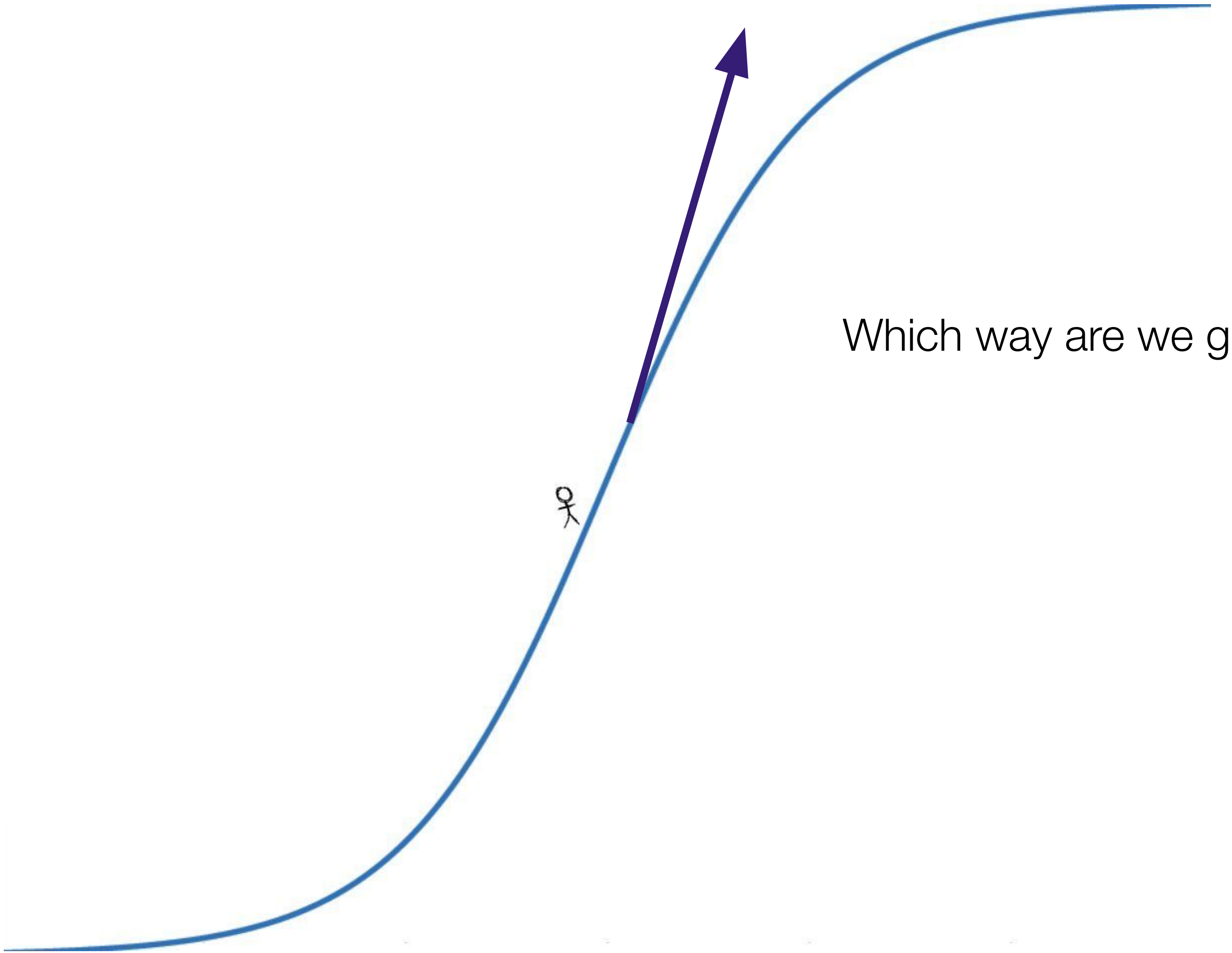
"The AI is a fun toy."

"The AI is helping me in some tasks."

"The AI has a jagged frontier, sometimes it's amazing, sometimes it's dumb."

"The AI is unbelievably intelligent but for some reason it fails at X."





Which way are we going?

What “AI going well” means?

- World stays ~same but a little better?
- Solve poverty, sickness, death?
- Dyson spheres?
- USA is #1?
- Overthrowing capitalism?
- Humans remain in control?
- Benevolent AI rulers?

What can go wrong?

- AI weights can be stolen and used for bad purposes
- As it’s doing work, AI is going on the web, looking up documentation of packages, reading papers, etc.. – sees untrusted (adversarial) content.
- Tasks become far too complex for humans to verify – need scalable oversight / control / monitoring methods.
- AI might not be honest about shortcomings in its work – might aim to please– reward hacking.
- AI might have different long-term goals which it pursues covertly – scheming.

One alignment story

AI systems improve in length of time of tasks they can handle:

METR estimate: 4x / year

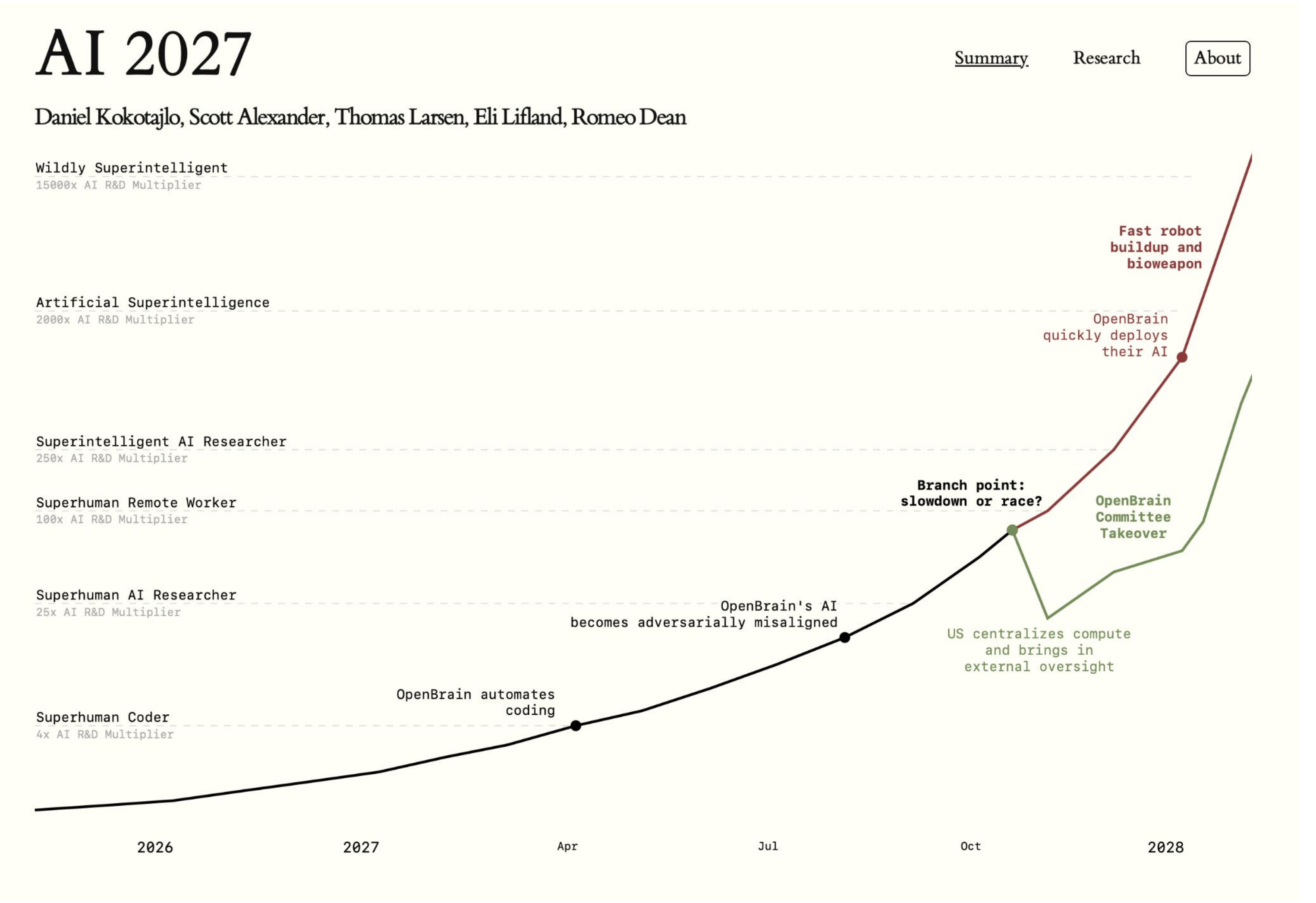
- ”Fix bug in this line of code”: ~30 sec
- “Write for me this function”: ~2m
- “Review this PR and fix bugs in it”: ~15m
- “Write this script for me and test it”: ~1h
- “Run this ML experiment for me”: ~4 hours
- “Check these N ablations for me and give me a report”: ~2 days
- “Suggest N ablations to improve efficiency and give me a report”~ 2 weeks
- …
- “Train your descendant” ~N person years

We are here

Potential Alignment Failures

- Sci-fi / genie failure 1: AI interprets instructions literally and result is opposite of what we want.
- Sci-fi failure 2: AI is power seeking / scheming
- Classic security failures: AI is ”hacked”/ “prompt injected”
- Classic security failure variant: “hack” is by another AI.
- Generalization failure: In very novel situation, AI generalizes badly.
- Alignment to humanity failure:
  - AI follows valid instructions, but these cause harm
  - Deploying AI at scale causes societal harm

AI 2027



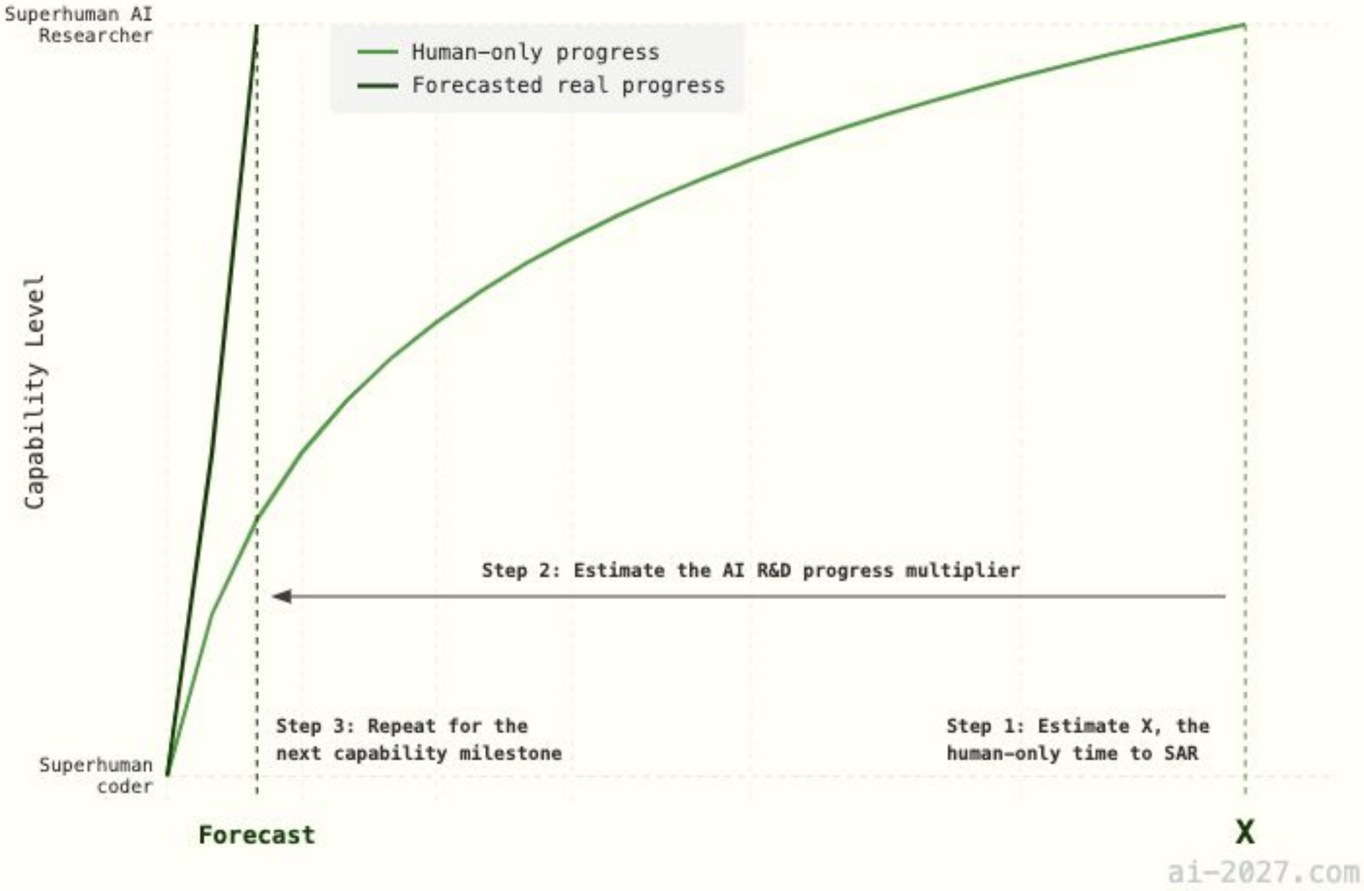
AI 2027

Summary Research About

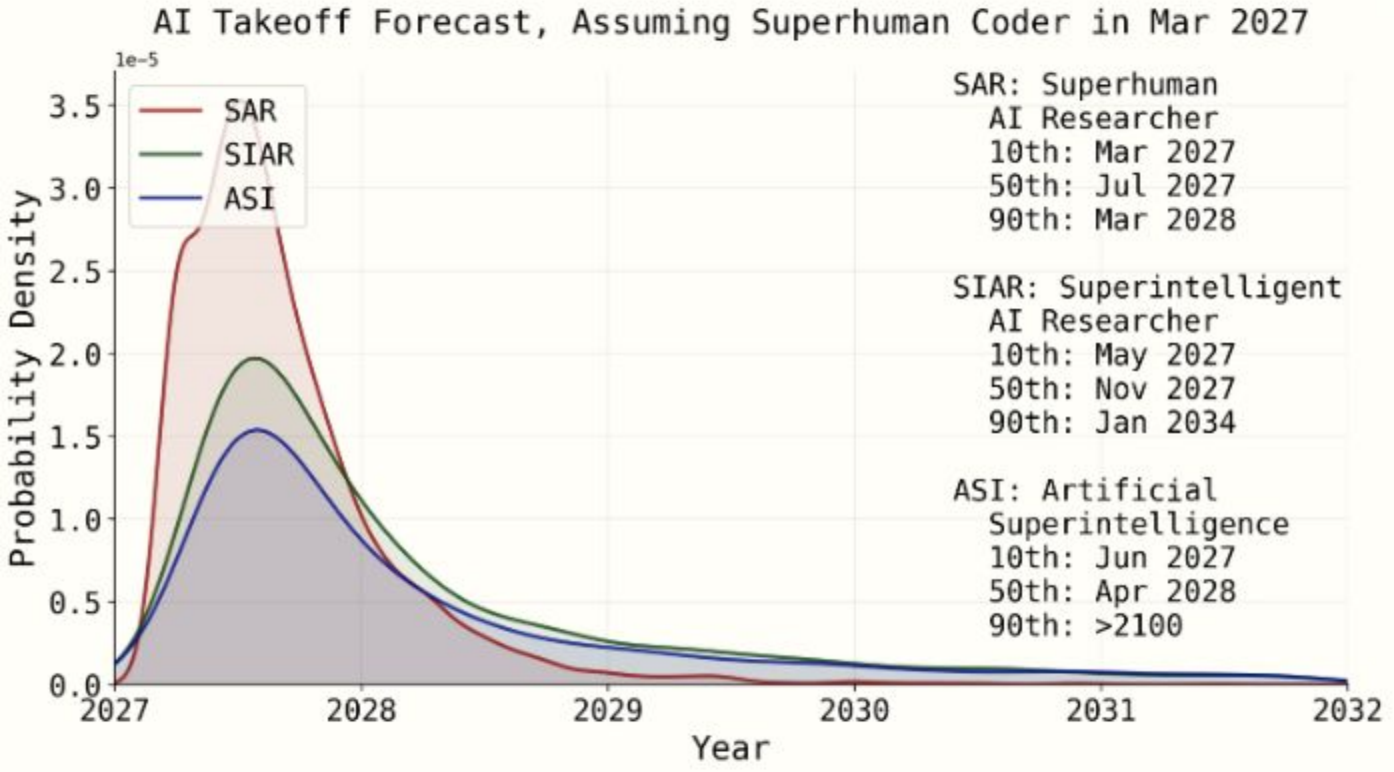
Daniel Kokotajlo, Scott Alexander, Thomas Larsen, Eli Lifland, Romeo Dean

Wildly Superintelligent
15000x AI R&D Multiplier
Artificial Superintelligence
2000x AI R&D Multiplier
Superintelligent AI Researcher
250x AI R&D Multiplier
Superhuman Remote Worker
100x AI R&D Multiplier
Superhuman AI Researcher
25x AI R&D Multiplier
Superhuman Coder
4x AI R&D Multiplier
Fast robot buildup and bioweapon
OpenBrain quickly deploys their AI
Branch point: slowdown or race?
OpenBrain Committee Takeover
OpenBrain's AI becomes adversarially misaligned
US centralizes compute and brings in external oversight
OpenBrain automates coding
2026
2027
Apr
Jul
Oct
2028

Takeoff Methodology


Superhuman AI Researcher
Human-only progress
Forecasted real progress
Capability Level
Step 2: Estimate the AI R&D progress multiplier
Step 3: Repeat for the next capability milestone
Step 1: Estimate X, the human-only time to SAR
Superhuman coder
Forecast
X
ai-2027.com

Milestone	Projected date conditional on SC in Mar 2027 (median + 80% CI)	Date achieved in scenario, racing ending[4]	Human-only, software-only time until next milestone (median + 80% CI)	AI R&D progress multiplier: Algorithmic progress speedup from AI vs. humans-only



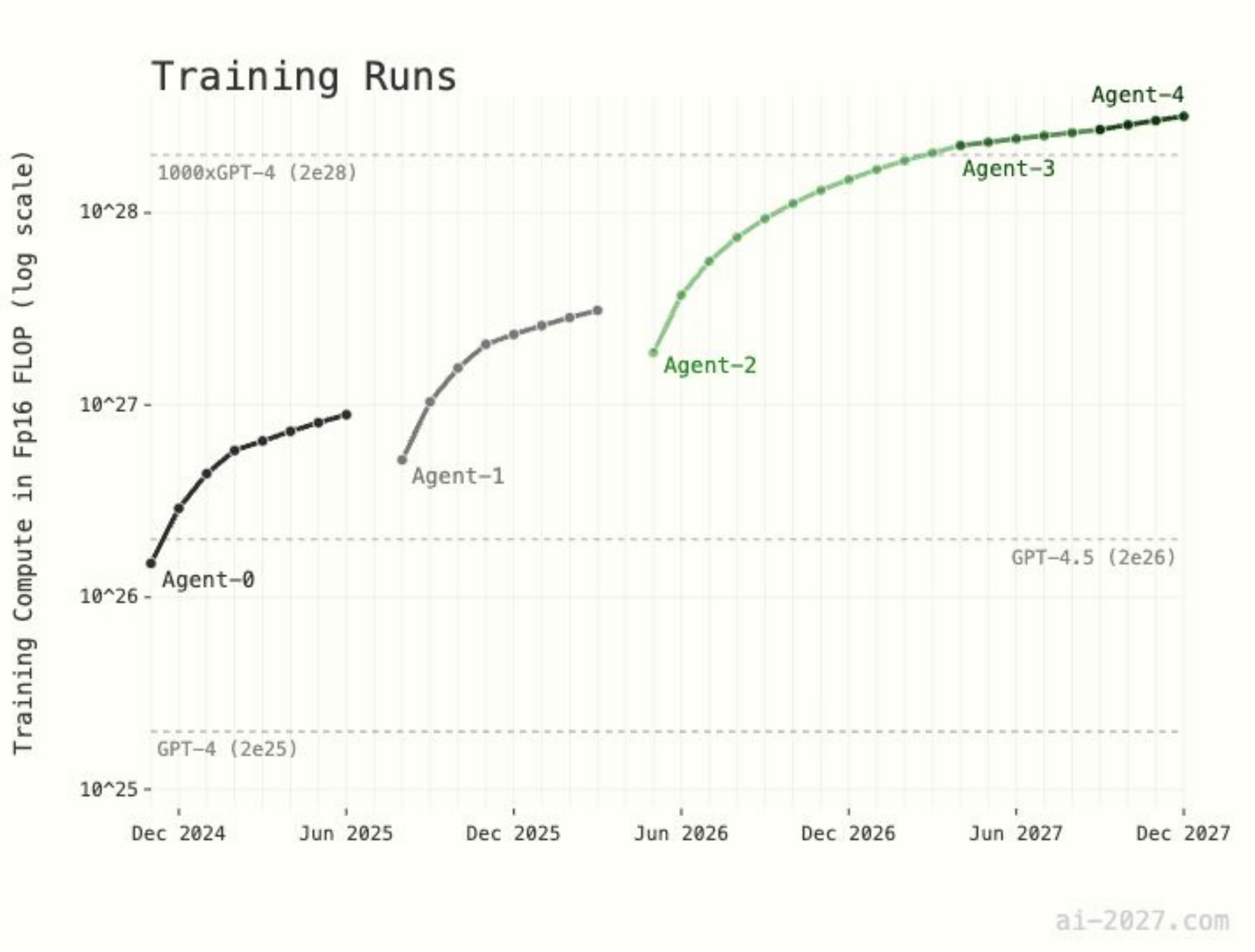
AI Takeoff Forecast, Assuming Superhuman Coder in Mar 2027

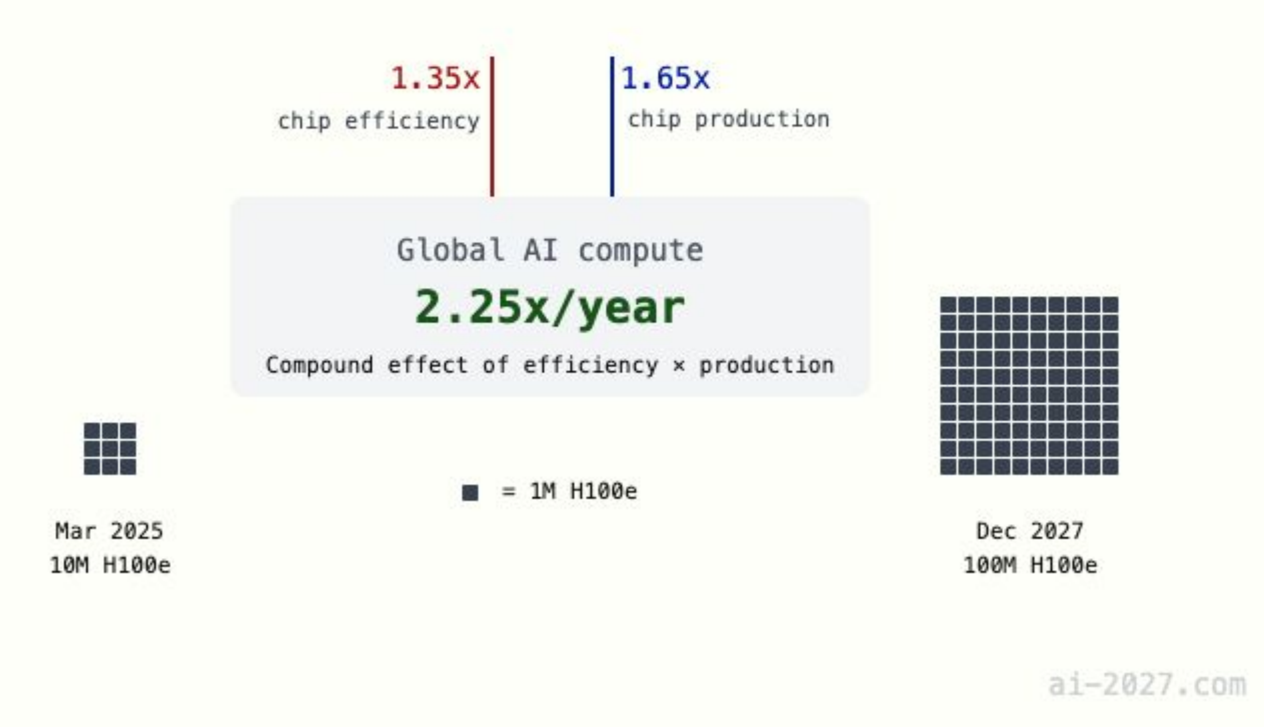
SAR: Superhuman
AI Researcher
10th: Mar 2027
50th: Jul 2027
90th: Mar 2028

SIAR: Superintelligent
AI Researcher
10th: May 2027
50th: Nov 2027
90th: Jan 2034

ASI: Artificial
Superintelligence
10th: Jun 2027
50th: Apr 2028
90th: >2100

ai-2027.com


Training Runs
Training Compute in Fp16 FLOP (log scale)
10^28
10^27
10^26
10^25
1000xGPT-4 (2e28)
GPT-4.5 (2e26)
GPT-4 (2e25)
Agent-0
Agent-1
Agent-2
Agent-3
Agent-4
Dec 2024
Jun 2025
Dec 2025
Jun 2026
Dec 2026
Jun 2027
Dec 2027
ai-2027.com



1.35x
chip efficiency

1.65x
chip production

Global AI compute

2.25x/year

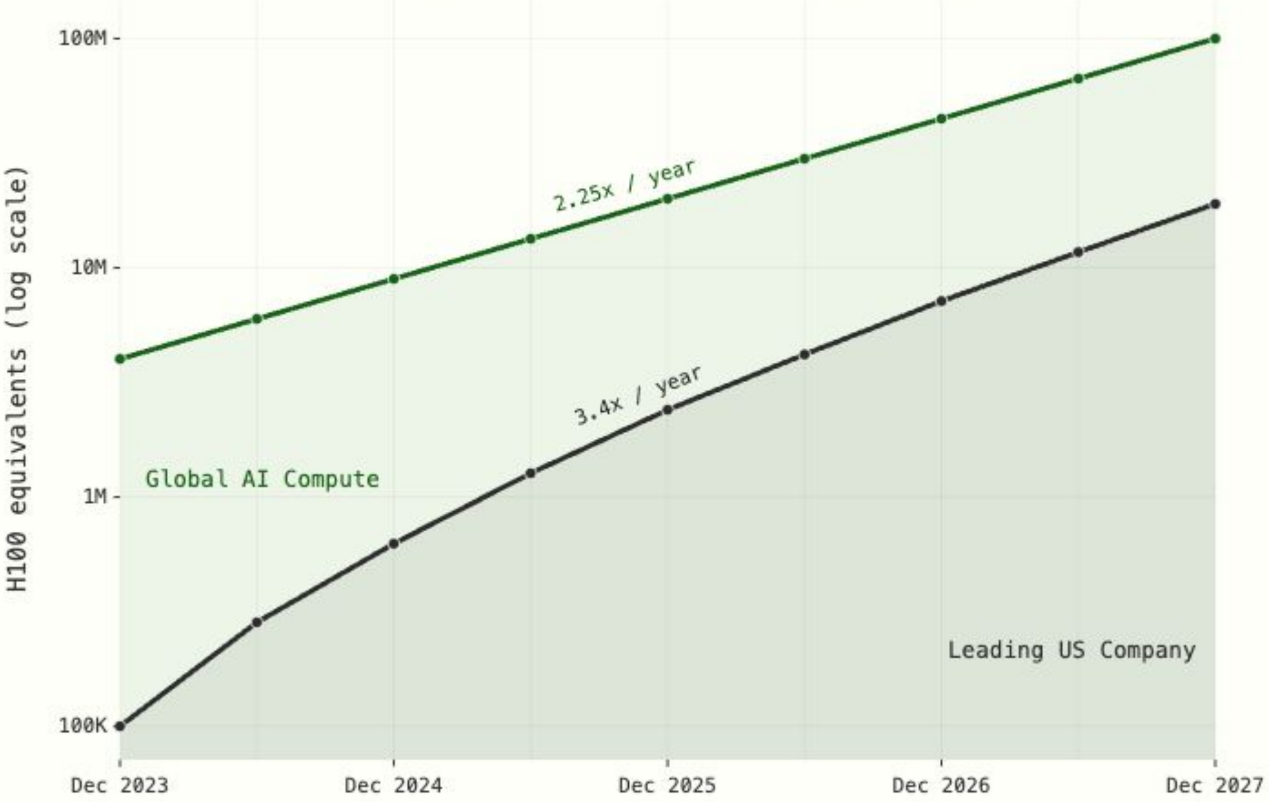
Compound effect of efficiency × production

■ = 1M H100e

Mar 2025
10M H100e

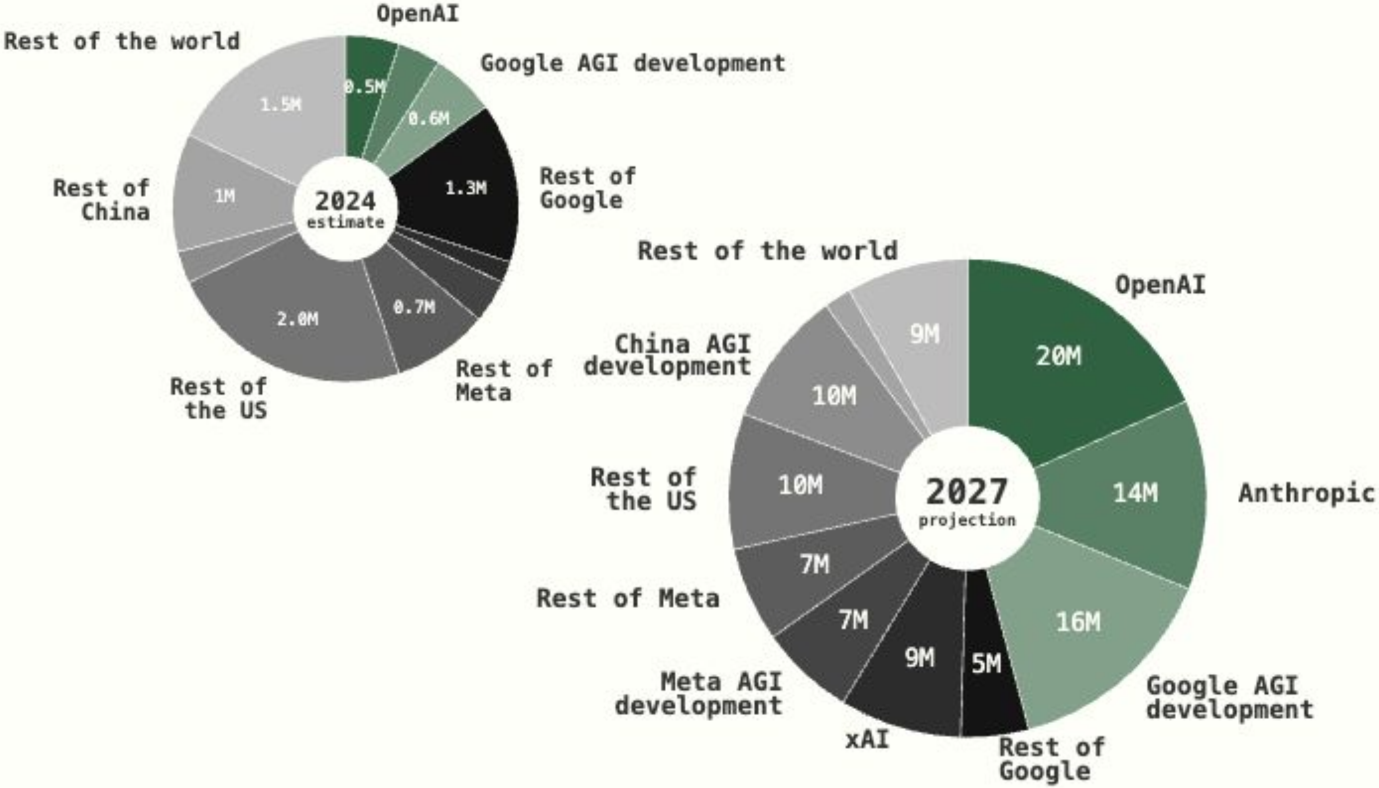
Dec 2027
100M H100e

Global AI Compute with Leading Company Share



ai-2027.com

Compute Breakdown by End-User, H100e



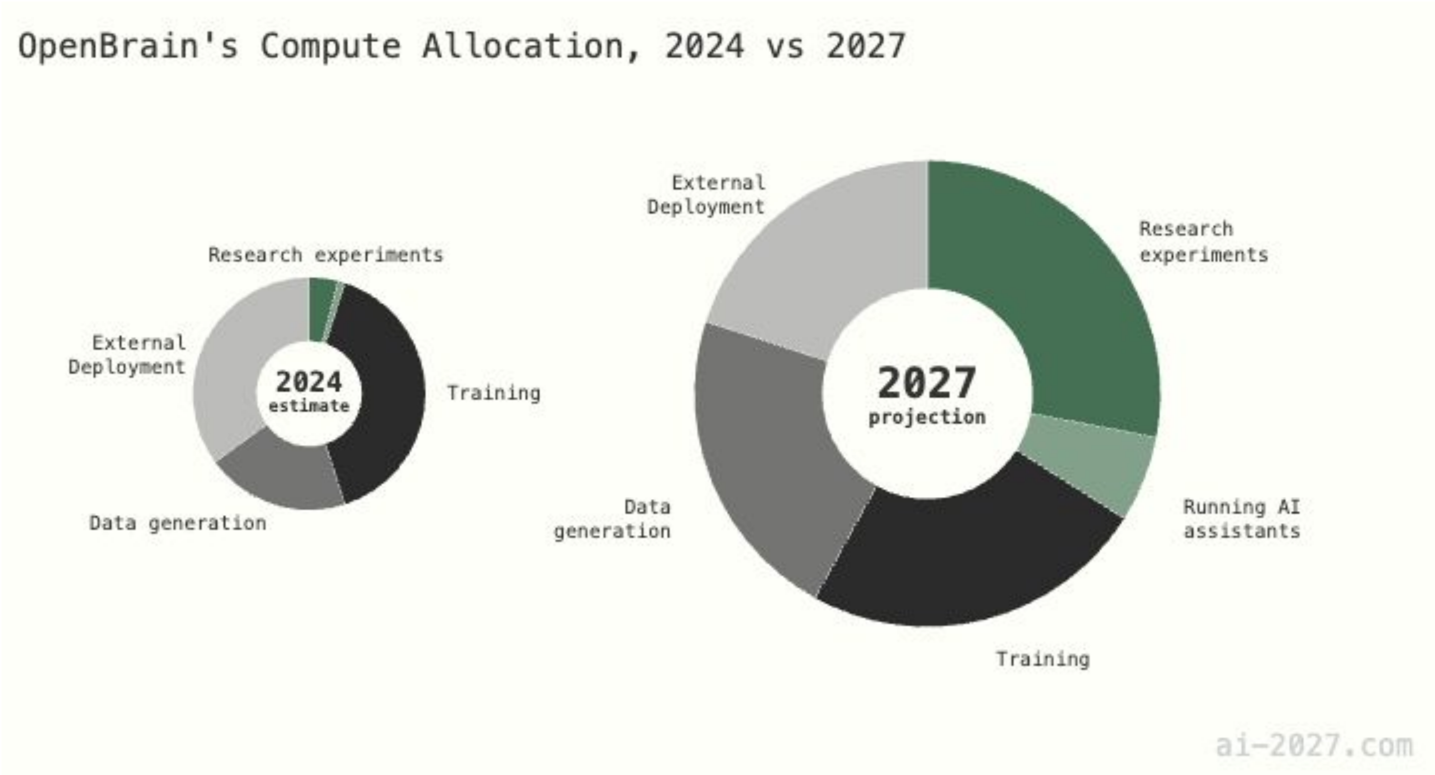
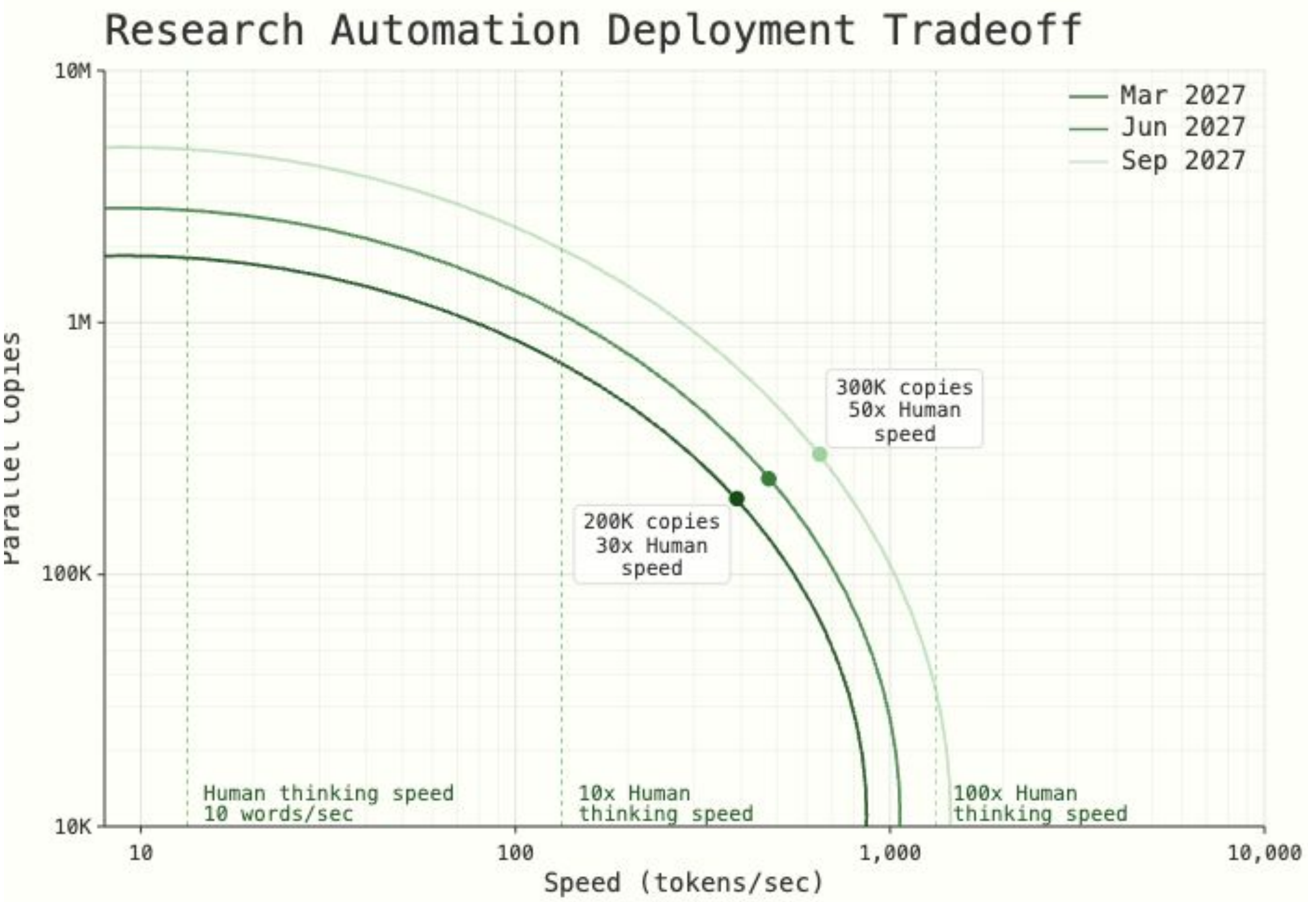
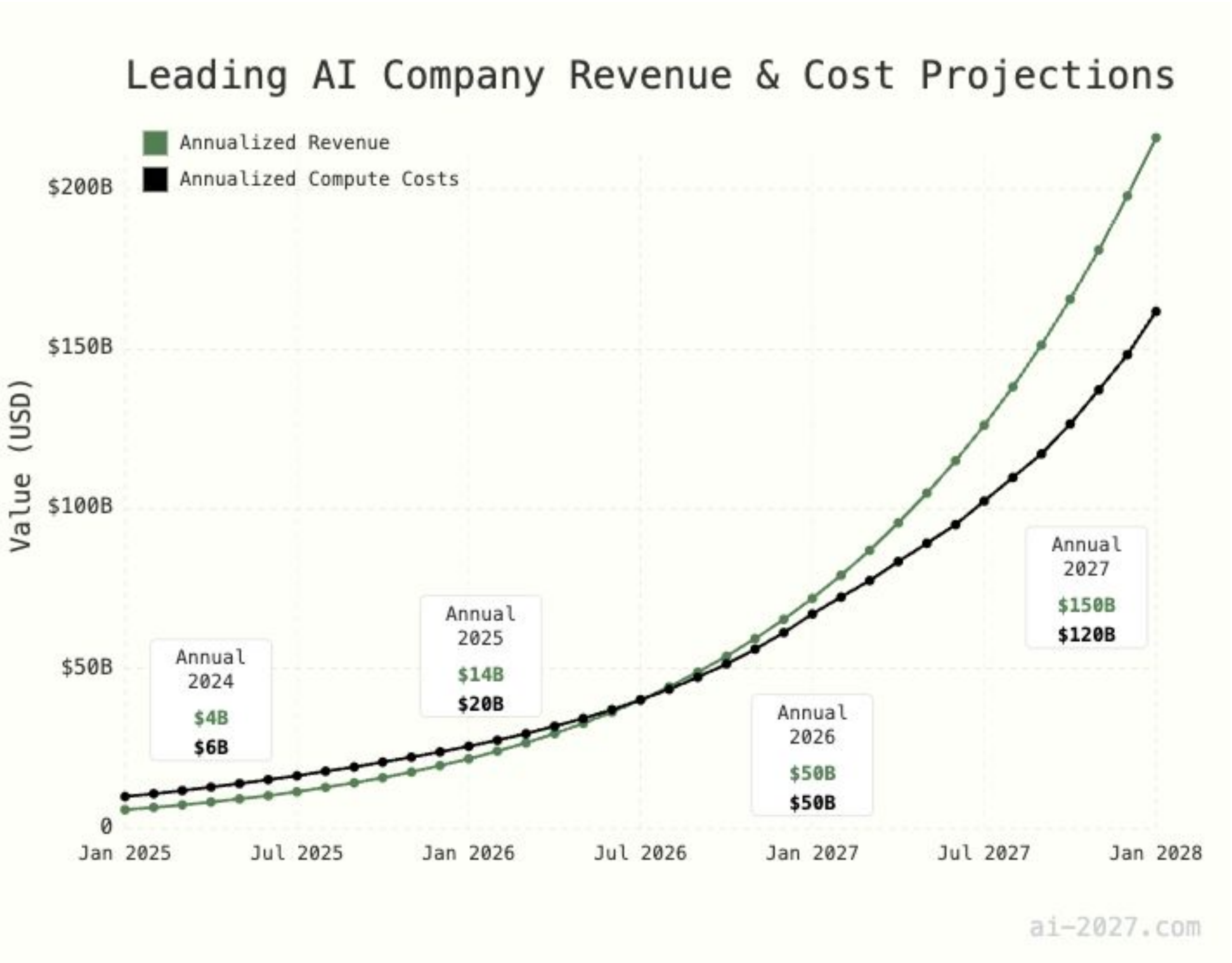
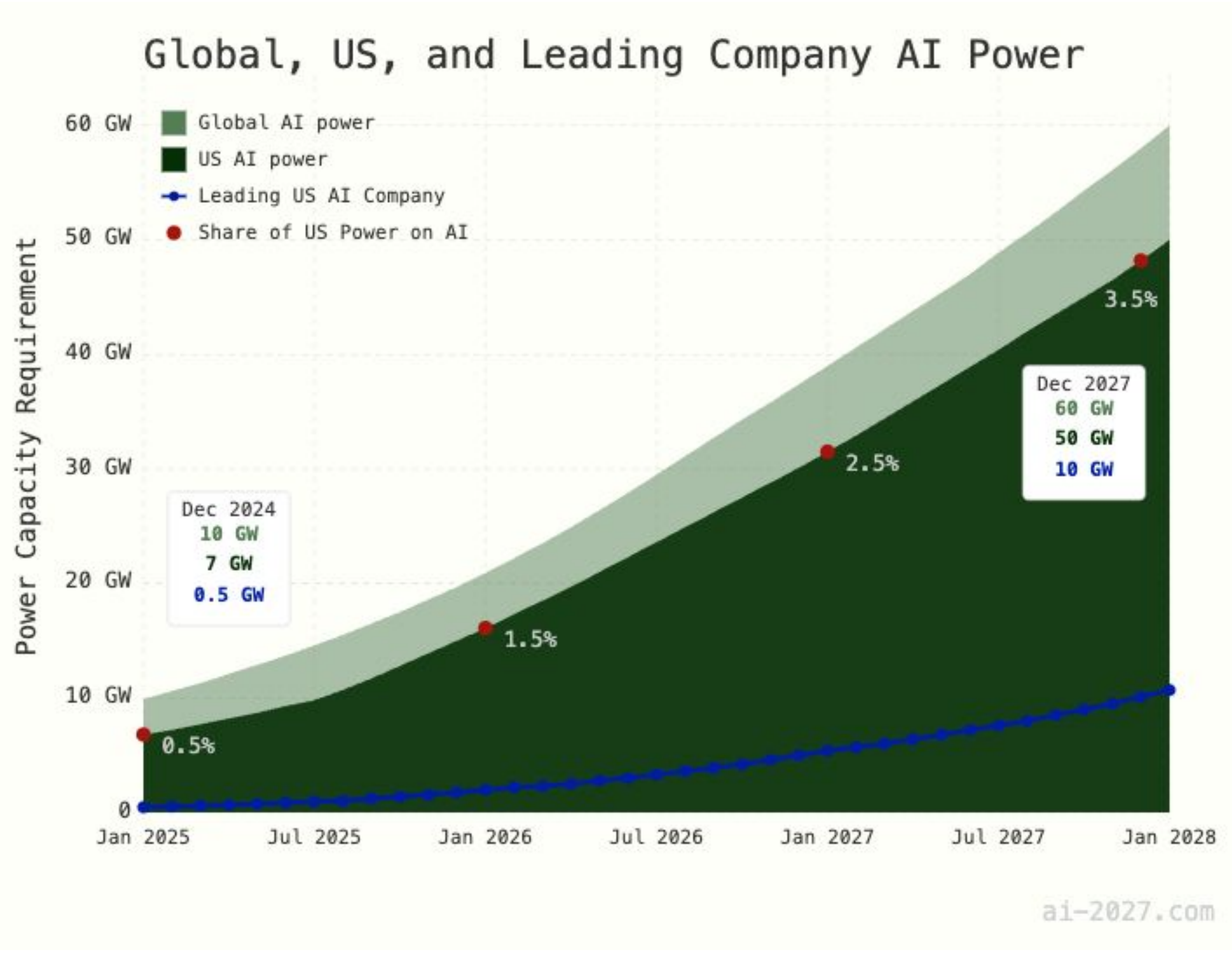


Figure 6: Compute use concentrates towards research automation. Data generation also increases.


Research Automation Deployment Tradeoff
Mar 2027
Jun 2027
Sep 2027
300K copies 50x Human speed
200K copies 30x Human speed
Human thinking speed 10 words/sec
10x Human thinking speed
100x Human thinking speed
Parallel Copies
10M
1M
100K
10K
10
100
1,000
10,000
Speed (tokens/sec)



Leading AI Company Revenue & Cost Projections
🟩 Annualized Revenue
■ Annualized Compute Costs
$200B
$150B
$100B
$50B
0
Value (USD)
Jan 2025
Jul 2025
Jan 2026
Jul 2026
Jan 2027
Jul 2027
Jan 2028
Annual 2024
$4B
$6B
Annual 2025
$14B
$20B
Annual 2026
$50B
$50B
Annual 2027
$150B
$120B
ai-2027.com



Global, US, and Leading Company AI Power

KEY METRICS 2027

GLOBAL AI CAPEX

$2T

COST OF OWNERSHIP OF ACTIVE COMPUTE

GLOBAL AI POWER

60GW

PEAK POWER

OPENBRAIN REVENUE

$140B

2027 ANNUAL

TOTAL CAPITAL EXPENDITURE

$400B

COST OF OWNERSHIP OF OPENBRAIN'S ACTIVE COMPUTE

SHARE OF US POWER ON AI

3.5%

50 GW OF 1.35TW CAPACITY

OPENBRAIN COMPUTE COSTS

$100B

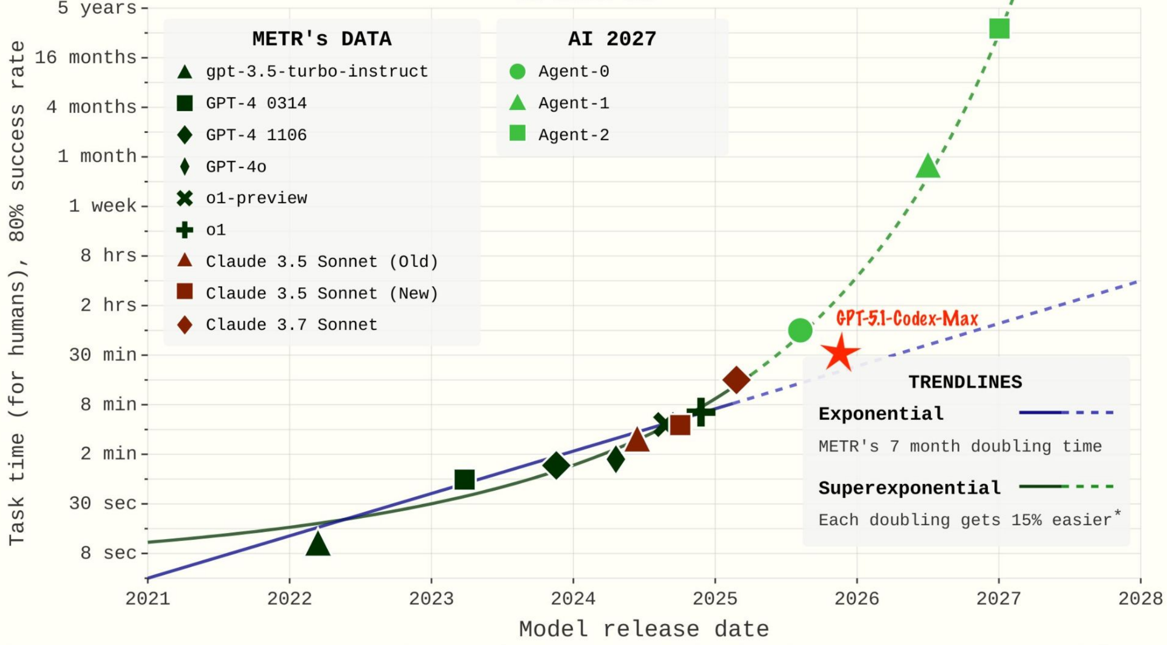
2027 ANNUAL

OPENBRAIN POWER REQUIREMENT

10GW

PEAK POWER

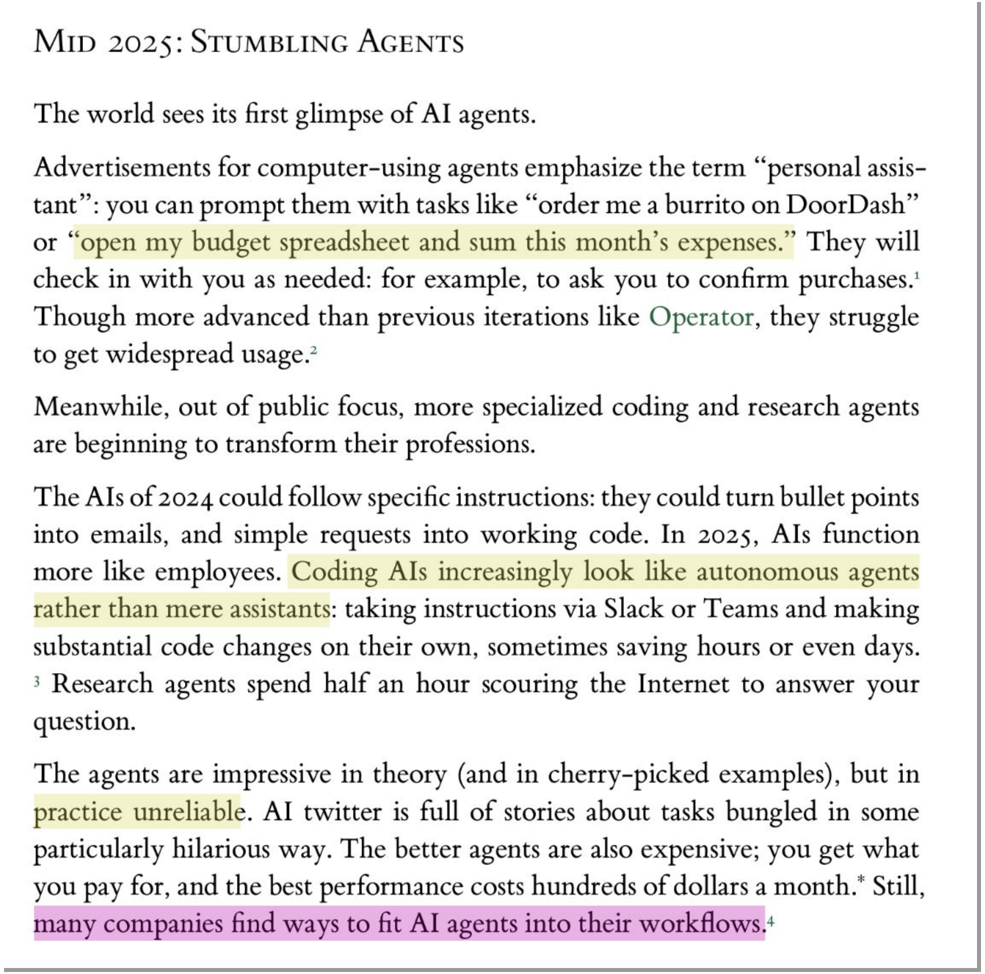
AI 2027
(critiques)



Length Of Coding Tasks AI Agents Can Complete Autonomously

80% success rate

Forecast: Doublings may get faster due to fewer new skills being needed at higher timescales, and automation of AI R&D. The green curve is essentially a simplified version of the full AI 2027 timelines model. Upon AI 2027 release, our full model did not "backcast" previous data points as well as this curve. As of Jul 2025, we're working on updates.



Mid 2025: Stumbling Agents

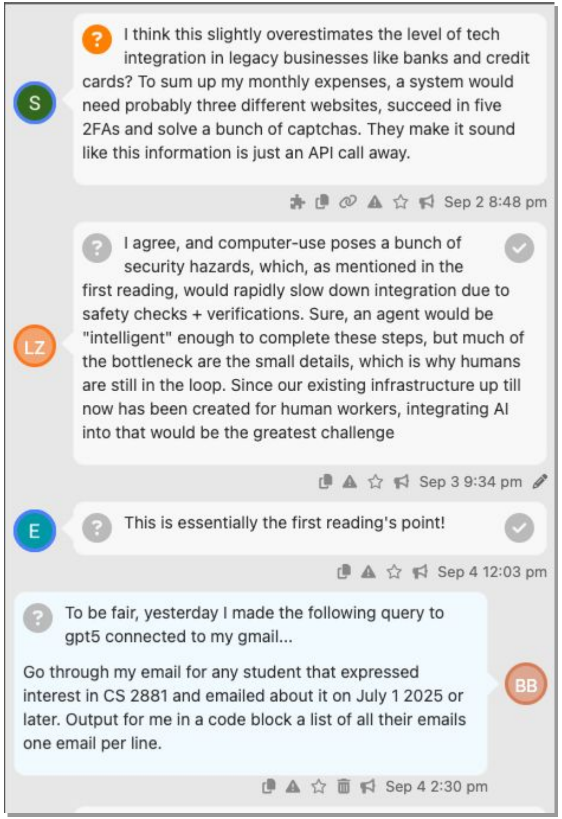
The world sees its first glimpse of AI agents.

Advertisements for computer-using agents emphasize the term "personal assistant": you can prompt them with tasks like "order me a burrito on DoorDash" or "open my budget spreadsheet and sum this month's expenses." They will check in with you as needed: for example, to ask you to confirm purchases.[1] Though more advanced than previous iterations like Operator, they struggle to get widespread usage.[2]

Meanwhile, out of public focus, more specialized coding and research agents are beginning to transform their professions.

The AIs of 2024 could follow specific instructions: they could turn bullet points into emails, and simple requests into working code. In 2025, AIs function more like employees. Coding AIs increasingly look like autonomous agents rather than mere assistants: taking instructions via Slack or Teams and making substantial code changes on their own, sometimes saving hours or even days. [3] Research agents spend half an hour scouring the Internet to answer your question.

The agents are impressive in theory (and in cherry-picked examples), but in practice unreliable. AI twitter is full of stories about tasks bungled in some particularly hilarious way. The better agents are also expensive; you get what you pay for, and the best performance costs hundreds of dollars a month.[*] Still, many companies find ways to fit AI agents into their workflows.[4]



S

? I think this slightly overestimates the level of tech integration in legacy businesses like banks and credit cards? To sum up my monthly expenses, a system would need probably three different websites, succeed in five 2FAs and solve a bunch of captchas. They make it sound like this information is just an API call away.

Sep 2 8:48 pm

LZ

? I agree, and computer-use poses a bunch of security hazards, which, as mentioned in the first reading, would rapidly slow down integration due to safety checks + verifications. Sure, an agent would be "intelligent" enough to complete these steps, but much of the bottleneck are the small details, which is why humans are still in the loop. Since our existing infrastructure up till now has been created for human workers, integrating AI into that would be the greatest challenge

Sep 3 9:34 pm

E

? This is essentially the first reading's point!

Sep 4 12:03 pm

BB

? To be fair, yesterday I made the following query to gpt5 connected to my gmail...

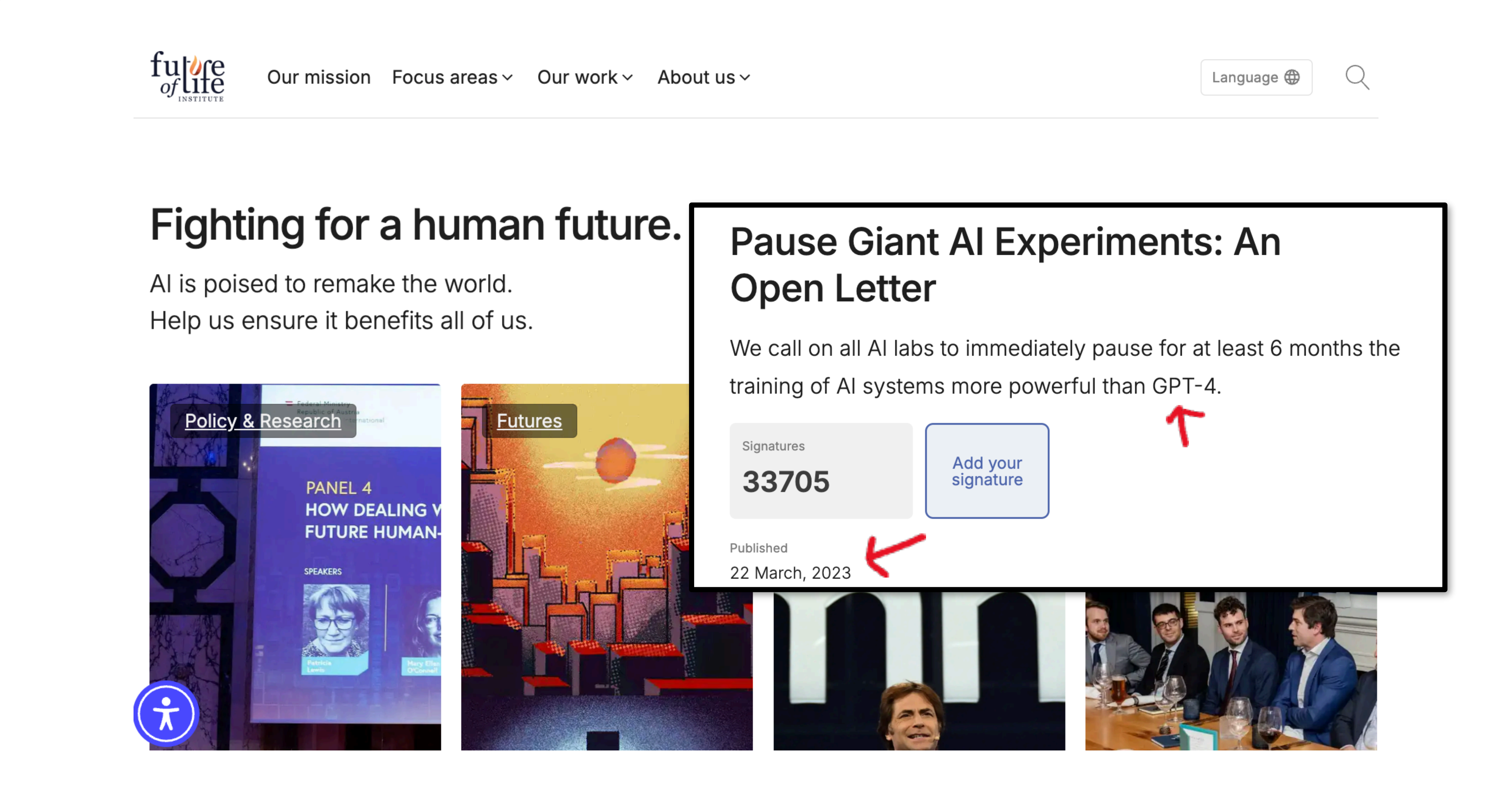
Go through my email for any student that expressed interest in CS 2881 and emailed about it on July 1 2025 or later. Output for me in a code block a list of all their emails one email per line.

Sep 4 2:30 pm

AI GOVERNANCE

VITALY SHMATIKOV

Lighthaven is a space dedicated to hosting events and programs that help people think better and to improve humanity's long-term trajectory.



Fighting for a human future.

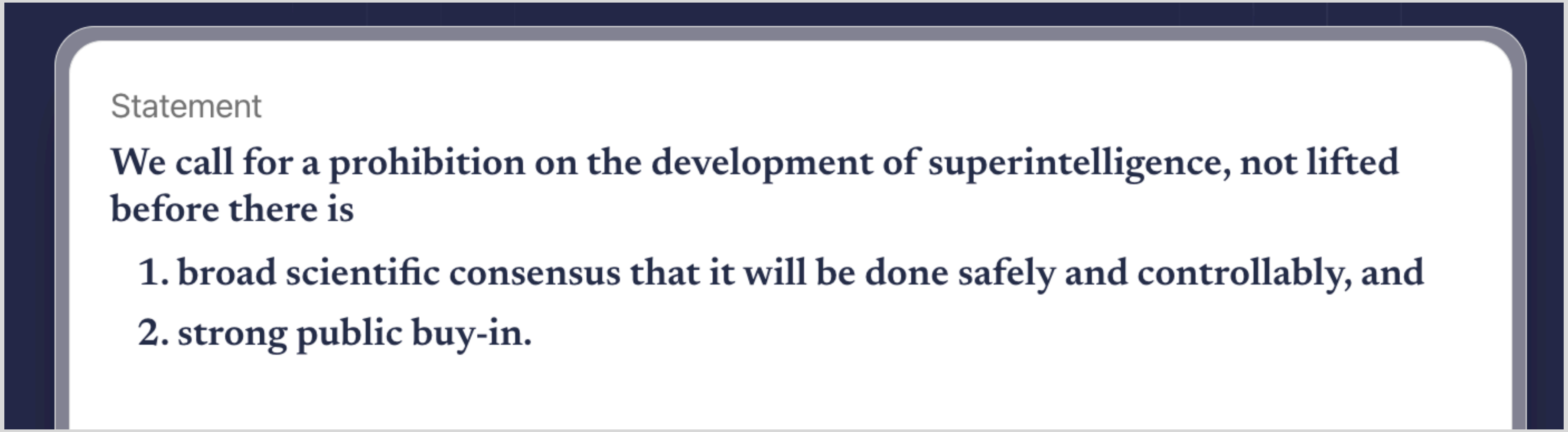
AI is poised to remake the world.
Help us ensure it benefits all of us.

Pause Giant AI Experiments: An Open Letter

We call on all AI labs to immediately pause for at least 6 months the training of AI systems more powerful than GPT-4.

Signatures
33705

Published
22 March, 2023



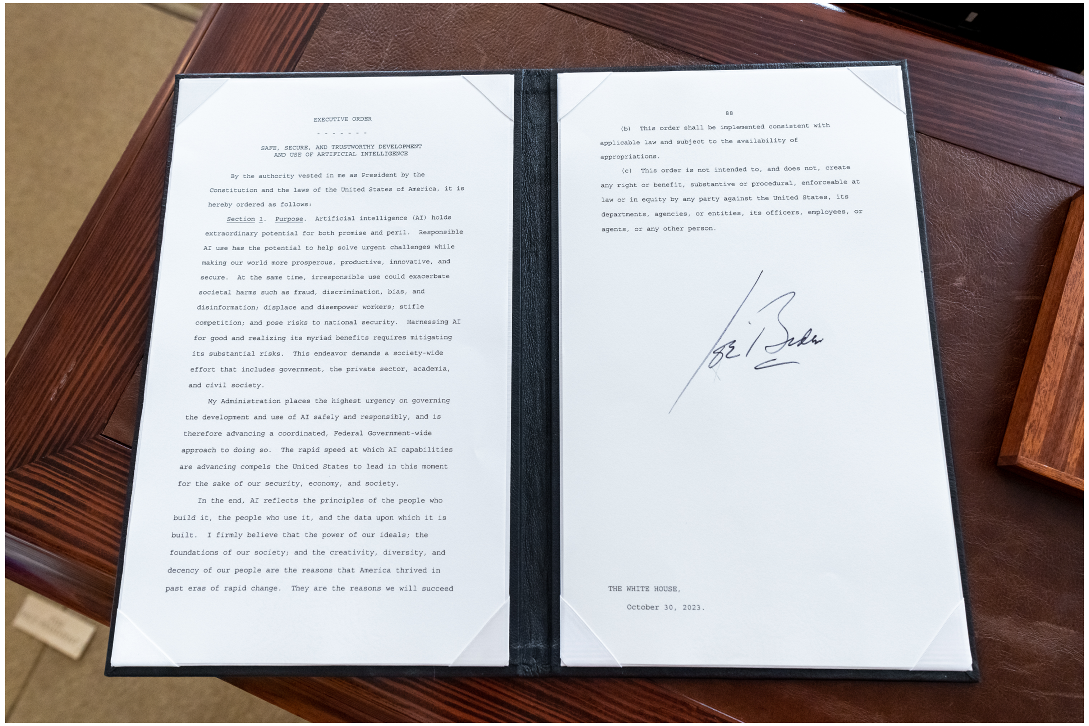
Statement

We call for a prohibition on the development of superintelligence, not lifted before there is

1. **broad scientific consensus that it will be done safely and controllably, and**
2. **strong public buy-in.**

https://superintelligence-statement.org/

Geoffrey Hinton, Yoshua Bengio, Stuart Russell, Steve Wozniak…
Steve Bannon, Susan Rice, Adm. Mike Mullen…
Prince Harry, Meghan Markle…



EXECUTIVE ORDER

- - - - - - - -

SAFE, SECURE, AND TRUSTWORTHY DEVELOPMENT
AND USE OF ARTIFICIAL INTELLIGENCE

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

Section 1. Purpose. Artificial intelligence (AI) holds extraordinary potential for both promise and peril. Responsible AI use has the potential to help solve urgent challenges while making our world more prosperous, productive, innovative, and secure. At the same time, irresponsible use could exacerbate societal harms such as fraud, discrimination, bias, and disinformation; displace and disempower workers; stifle competition; and pose risks to national security. Harnessing AI for good and realizing its myriad benefits requires mitigating its substantial risks. This endeavor demands a society-wide effort that includes government, the private sector, academia, and civil society.

My Administration places the highest urgency on governing the development and use of AI safely and responsibly, and is therefore advancing a coordinated, Federal Government-wide approach to doing so. The rapid speed at which AI capabilities are advancing compels the United States to lead in this moment for the sake of our security, economy, and society.

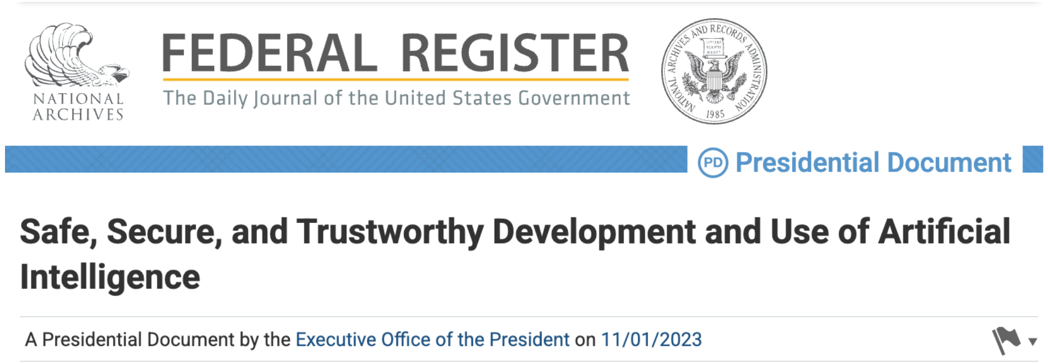
In the end, AI reflects the principles of the people who build it, the people who use it, and the data upon which it is built. I firmly believe that the power of our ideals; the foundations of our society; and the creativity, diversity, and decency of our people are the reasons that America thrived in past eras of rapid change. They are the reasons we will succeed

88

(b) This order shall be implemented consistent with applicable law and subject to the availability of appropriations.

(c) This order is not intended to, and does not, create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person.

THE WHITE HOUSE,
October 30, 2023.



NATIONAL ARCHIVES

FEDERAL REGISTER
The Daily Journal of the United States Government

PD Presidential Document

Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

A Presidential Document by the Executive Office of the President on 11/01/2023

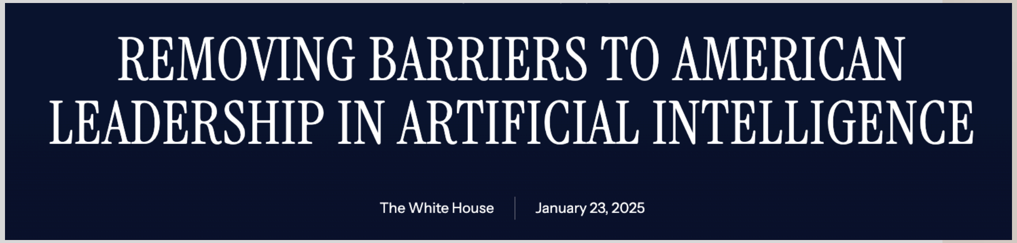
EO 14110

Calls for evaluations, reports, frameworks, guidelines, and best practices for the development and deployment of “safe, secure, and trustworthy AI systems”

Mostly focuses on federal agencies

Private companies are directed to share with the federal government the results of “red-team” safety tests of foundation models

EO 14179

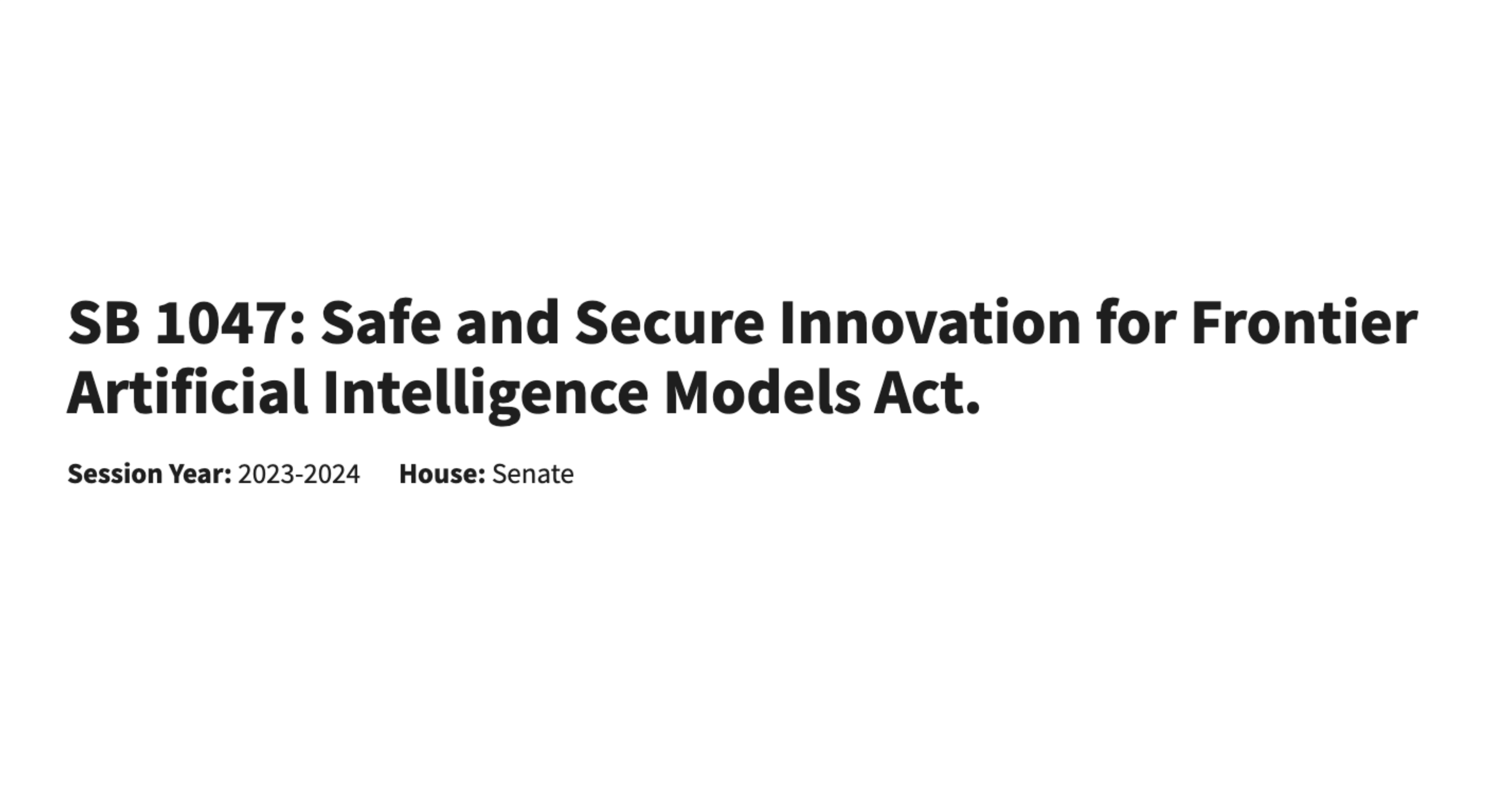


REMOVING BARRIERS TO AMERICAN LEADERSHIP IN ARTIFICIAL INTELLIGENCE

The White House | January 23, 2025

Revokes Biden's EO 14410

"dangerous… unnecessarily burdensome…"



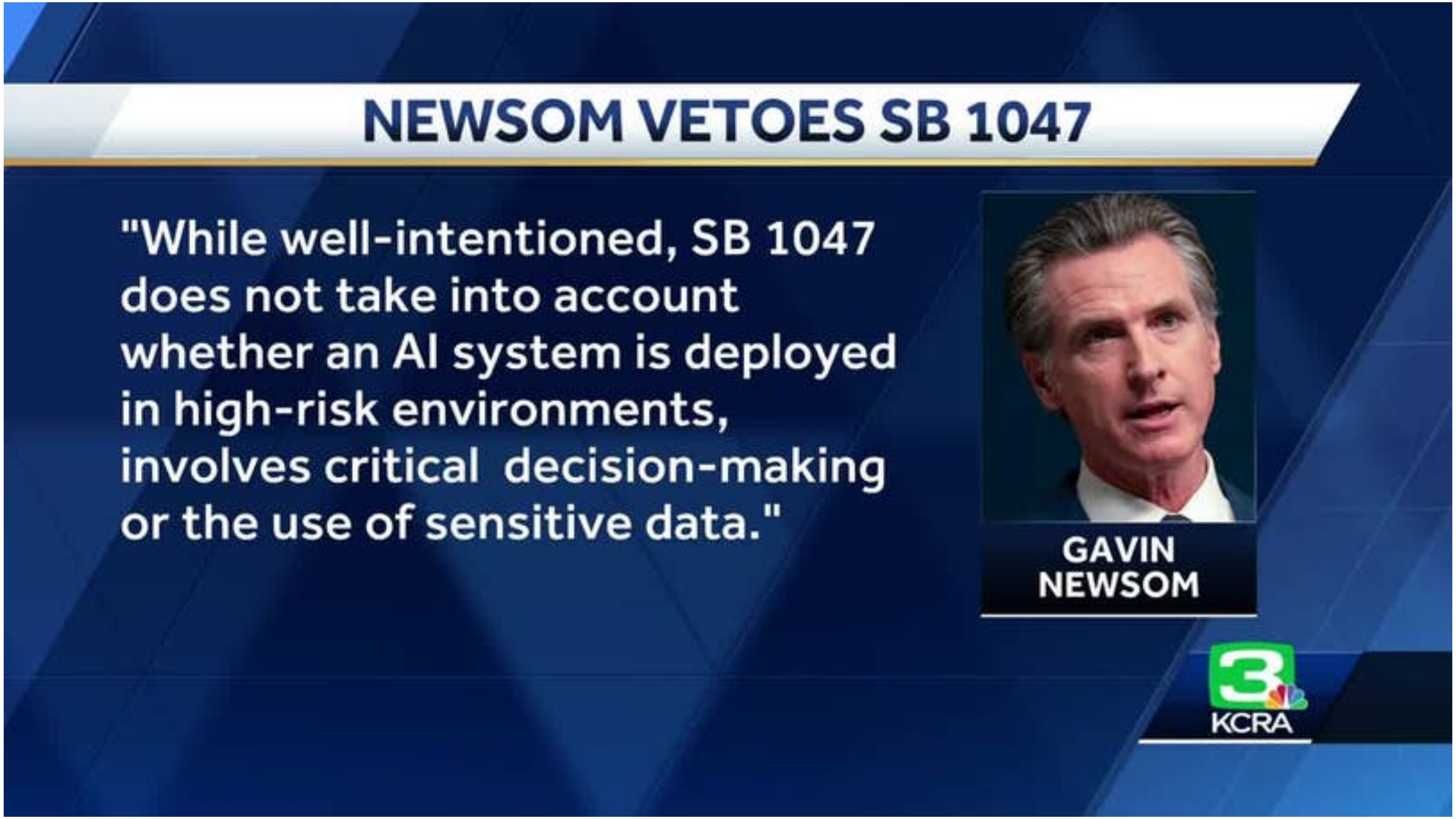
SB 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.

Session Year: 2023-2024 **House:** Senate

California SB 1047

Covers models costing $100+ million to train or $10+ million to fine-tur

- Comprehensive, documented safety and security measures for training, protections from misuse, unauthorized access, unsafe alterations
- “Kill switches” / mandatory shutdown capabilities
- Developers must ensure users cannot cause “critical harms” (chemical, bio, nuclear weapons, cyberattacks, grave physical damage)
- Audits, certification, annual reporting with severe penalties for non-compliance
- Stringent KYC and shutdown requirements on operators of compute clusters

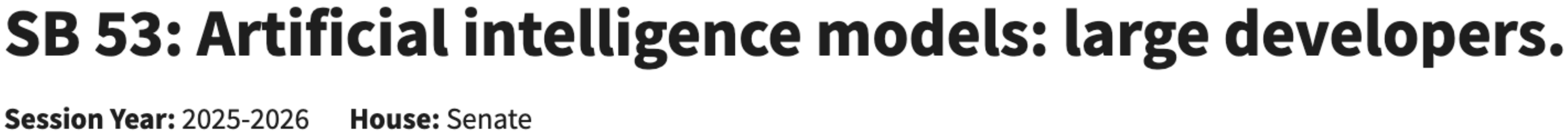


NEWSOM VETOES SB 1047

"While well-intentioned, SB 1047 does not take into account whether an AI system is deployed in high-risk environments, involves critical decision-making or the use of sensitive data."

GAVIN NEWSOM

https://www.kcra.com/article/california-governor-vetoes-bill-ai-safety-measures/62436875



SB 53: Artificial intelligence models: large developers.

Session Year: 2025-2026 **House:** Senate

TFAIA: Transparency in Frontier Artificial Intelligence Act

California SB 53

Covers models trained with >10^{26} FLOPs

- Requires enterprise frameworks for identifying and mitigating catastrophic risks and securing unreleased model weights
- Requires **transparency**: reporting of model capabilities, intended uses, risk assessments, third-party evaluations, and mitigation measures
  - ”Large frontier developers” ($500M+ revenue) must publish how they incorporate standards and industry-consensus best practices
- Requires reporting of safety incidents
- ~~Pre-deployment testing~~
- ~~Mandatory shutdown capabilities~~

Other State and Local AI Regulation

- California rules on “automated decision-making technology” (under California Consumer Privacy Act)
- Colorado AI antidiscrimination law
- Virginia AI antidiscrimination law (vetoed by previous governor)
- Illinois law requiring notice of AI use in workplace
- New York City AI bias audit law

States’ regulatory frenzy is encouraging woke AI and will lead to a patchwork of laws that will undermine competitiveness of American AI industry

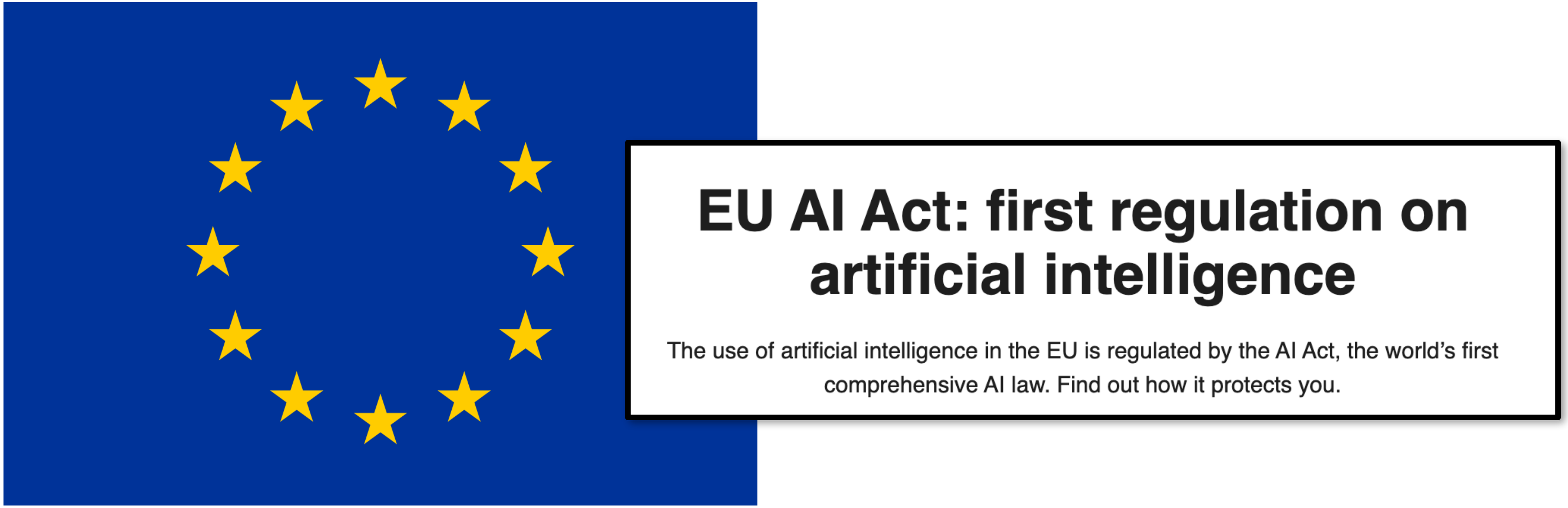
David Sacks
White House AI and Crypto Czar

Federal Preemption of AI Governance

As of 2025, White House, Republicans in Congress, and major AI companies want federal regulation to preempt state laws

- Amendments to major House bills to pause state-level regulation
- Attempts to add federal pre-emption language to NDAA (main defense funding bill)
- Planned (?) executive order

So far, all have failed



EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU is regulated by the AI Act, the world's first comprehensive AI law. Find out how it protects you.

Key Elements of EU AI Act

Categorization of AI systems by risk level:

- **Minimal**
- **General-purpose** – transparency, compliance with EU copyright law
- **High-risk**: healthcare, education, law enforcement, products that fall under EU product safety legislation – continuous assessment
- **Unacceptable:** behavioral manipulation, social scoring, exploitation of children, biometric identification and categorization – **banned**

Also requires model evaluations (incl. adversarial testing) to identify and mitigate systemic risks and tracking of serious incidents

References

1. https://ai-2027.com/
2. https://boazbk.github.io/mltheoryseminar/
3. https://nickbostrom.com/superintelligence
4. https://en.wikipedia.org/wiki/Moravec%27s_paradox