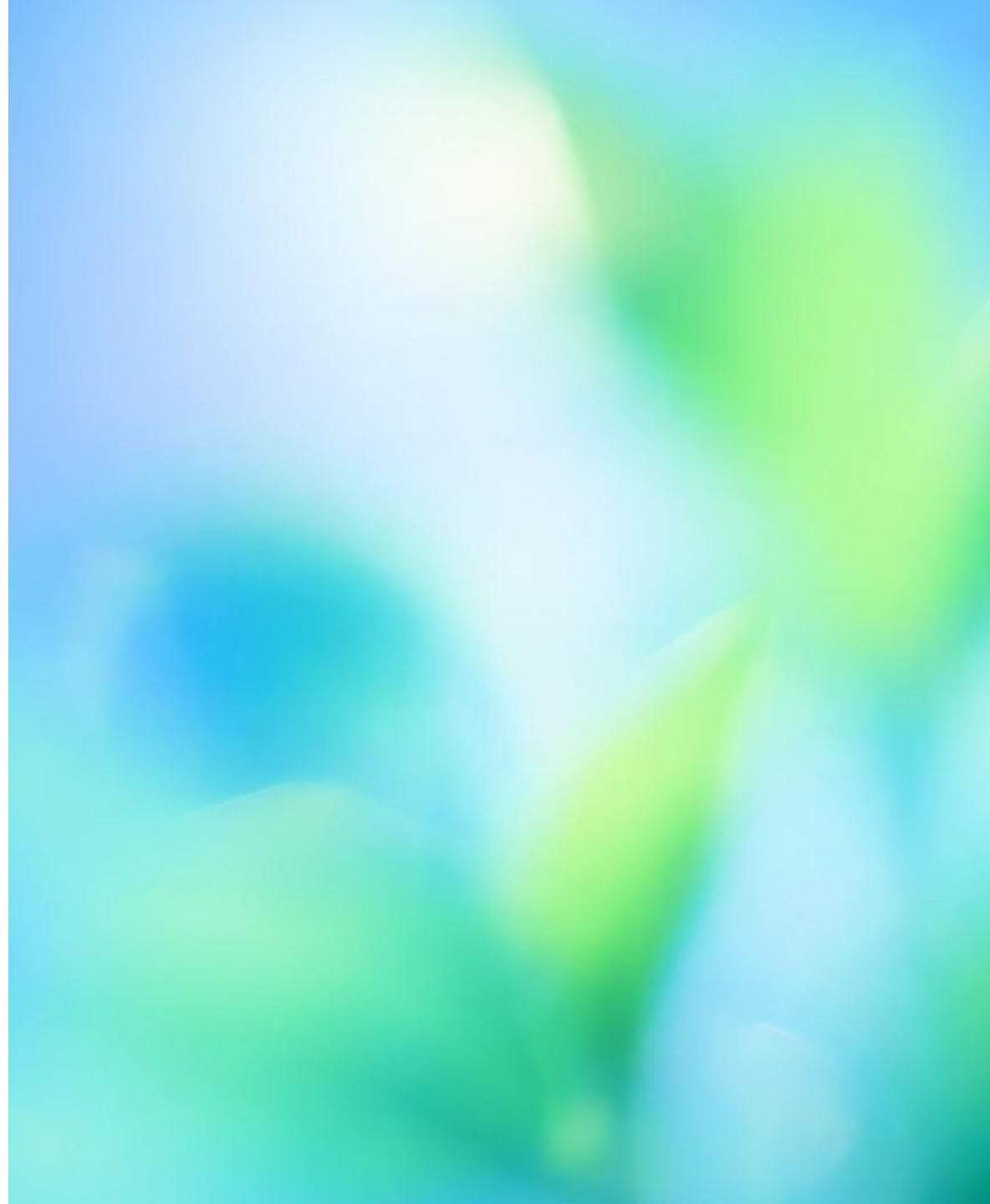


CS 1998

Forecasting, Policy, & Governance

Week 7

Introduction to AI Safety & Alignment



Logistics

- Project Proposal and Project Instructions up on Canvas!
 - They're due by EOD Oct 25
 - You can submit project proposal first to get feedback.

Course Evaluation

- This is REALLY important for us!
- Extra credit: 10 points
- The question we want to hear about is: *What could we have done better for the course? This can include anything, so feel free to share your thoughts on advertising, lectures, notebooks, projects, discussions, office hours, or support. We're running this course in future semesters, and we really appreciate your feedback!*

Spectrum!

"AI capability will hit a wall"

< A wall before 2030, from running out of data, compute, or ideas

Scaling will continue to get us superintelligent AI >

Spectrum!

"AI will bring transformative change at a scale equivalent or larger than the industrial revolution"

< AI is just a tool, like the internet or spreadsheets

AI will change everything forever >

Spectrum!

"Future models will be aligned by default"

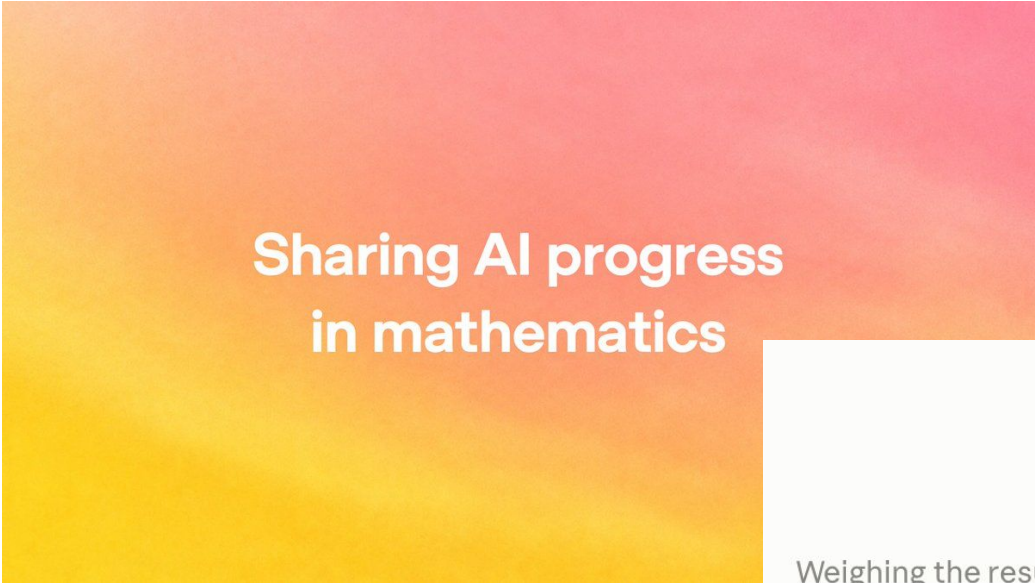
< Aligned by default: occasional misfires from misreading what people want, but it stays within human control

Uncontrollable by default: capable AI escapes human control unless we deliberately prevent it >

What do you think AI will be in 2 years?

- About 4 years ago, ChatGPT just came out
- 2 years ago, GPT-4o came out
- Today, AI can solve the Millenium problem

What do you think AI will be in 2 years?



Sharing AI progress
in mathematics

Assuming all proofs are correct, how big is this?

Weighing the results against history's biggest mathematical leaps. >

Assuming every proof is correct, this is the largest event in the history of mathematics. Nothing else comes close in breadth. My rough guess is that it amounts to several decades of top-tier progress across nearly every field, compressed into a few weeks.

What do you think AI will be in 2 years?

Share
in

For the most important results in this list, how big are they? If a single person comes up with a single result, would they win the Fields medal? Could a result be the flagship work of the career of a professor at Harvard/MIT?

Refining tiers with precedent cases from past prize decisions. >

Yes on both counts, for the top results. On the Harvard/MIT question, most of the headline results would clear that bar easily, and a handful would probably win a Fields Medal on their own.

How the Fields bar works. The medal goes to mathematicians under 40, usually for a body of work, but a single result can carry it. Roth won in 1958 essentially for proving that algebraic numbers have irrationality exponent 2. Viazovska's win rested largely on sphere packing. Committees also weigh the new ideas in a proof, not just the statement it proves. So the tiers below assume each proof introduces new methods proportionate to the problem.

Beyond a typical Fields

- **Quasi-RH:** For 70 years, nobody has proved a zero-free half-plane for $\zeta(s)$ for any $\delta > 0$. Proving $\Re s > 7/8$ for all Dirichlet L -functions, which also kills Siegel zeros, would be among the biggest number theory results of the century.
- **Hodge for CM abelian varieties:** This would bring with it the Tate conjecture for abelian varieties over finite fields and the Hodge standard conjecture for abelian varieties. These are central problems from the Grothendieck era.

Likely a Fields Medal by itself

- $\mu(\pi) = 2$: This is the direct analogue of Roth's theorem for a transcendental number. The previous best bound was around 7.1.
- **Erdős's arithmetic progression conjecture with quasipolynomial Szemerédi bounds:** Gowers's Fields Medal partly rested on much weaker bounds.
- **Goldfeld's conjecture plus the full BSD formula in corank ≤ 1 :** Bhargava's average-rank work was part of his Fields citation.
- **Hilbert's tenth problem over $\mathbb{Q}$:** This one is borderline to strong, depending on the year's



Assuming all proofs are correct, how big is this?

tical leaps. >

vent in the history of mathematics. Nothing
t amounts to several decades of top-tier
a few weeks.

What do you think AI will be in 2 years?

Today's Plan

Part 1: Forecasting

How fast is AI moving?

- 1 How we got here
- 2 Measuring the trend
- 3 What forecasters say
- 4 What it means for you

Part 2: Governance

Who gets to steer it?

- 1 Why govern now
- 2 Pacing the frontier
- 3 The policy toolkit
- 4 States, nations, and the world

Part 1

Forecasting

How fast is AI moving, and what should you do about it?

How we got here

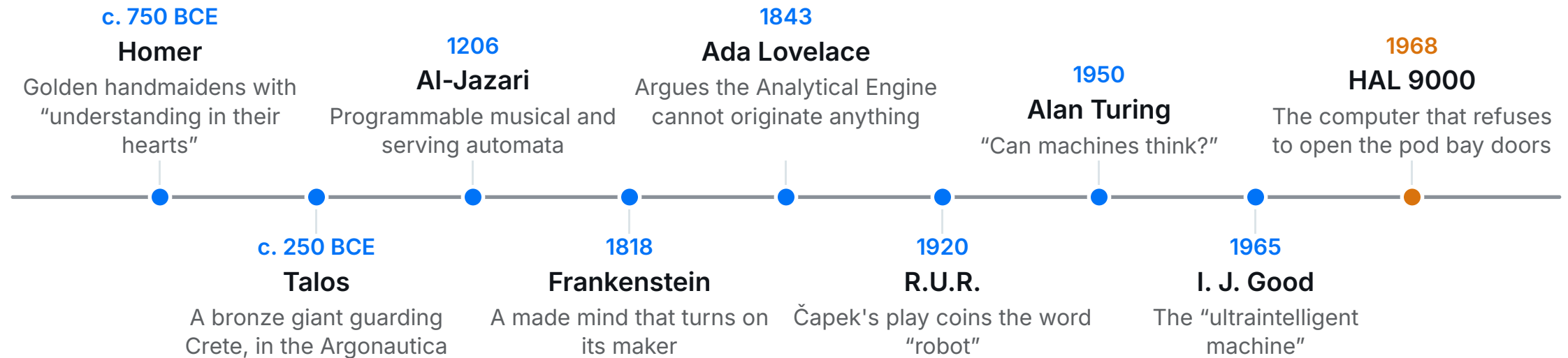
Measuring the trend

What forecasters say

What it means for you

From Homer to HAL

People imagined thinking machines for nearly 3,000 years before anyone built one.



The Hope and the Fear, Written Down in the 1960s

"The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

I. J. Good, Speculations Concerning the First Ultraintelligent Machine, 1965

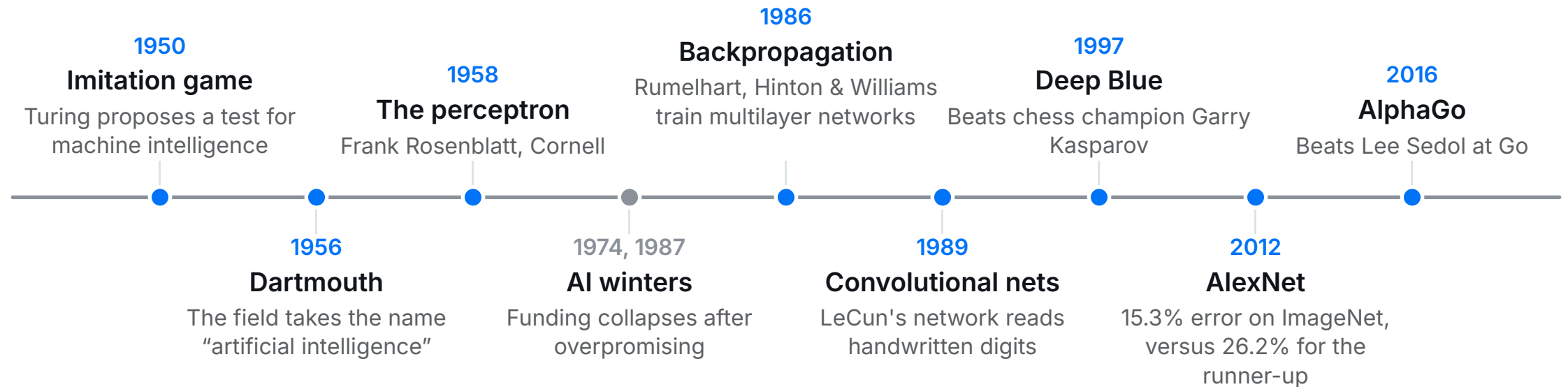
"I'm sorry, Dave. I'm afraid I can't do that."

HAL 9000, 2001: A Space Odyssey, 1968

The question for today is how close this is, and what to do about it.

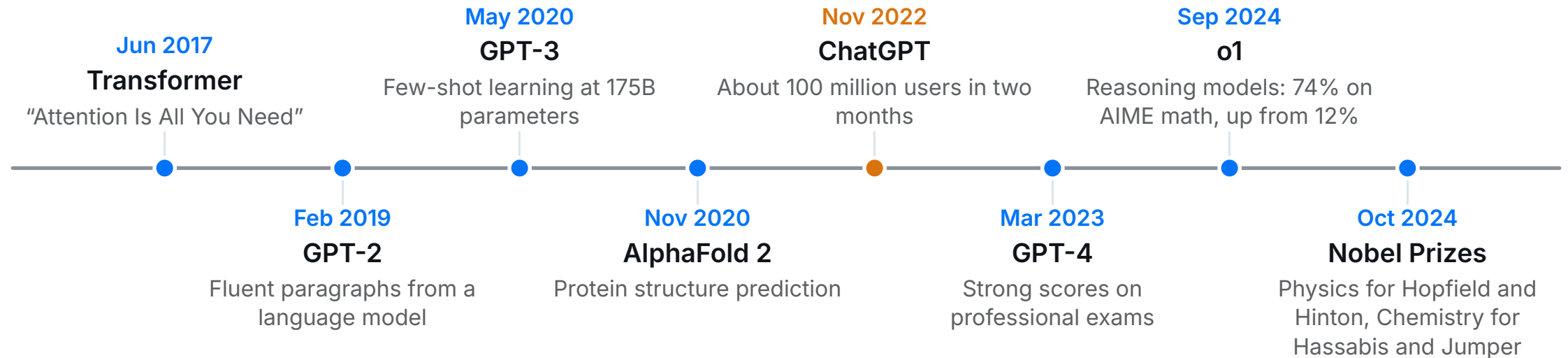
Seventy Years of Slow, Uneven Progress

1950 to 2016



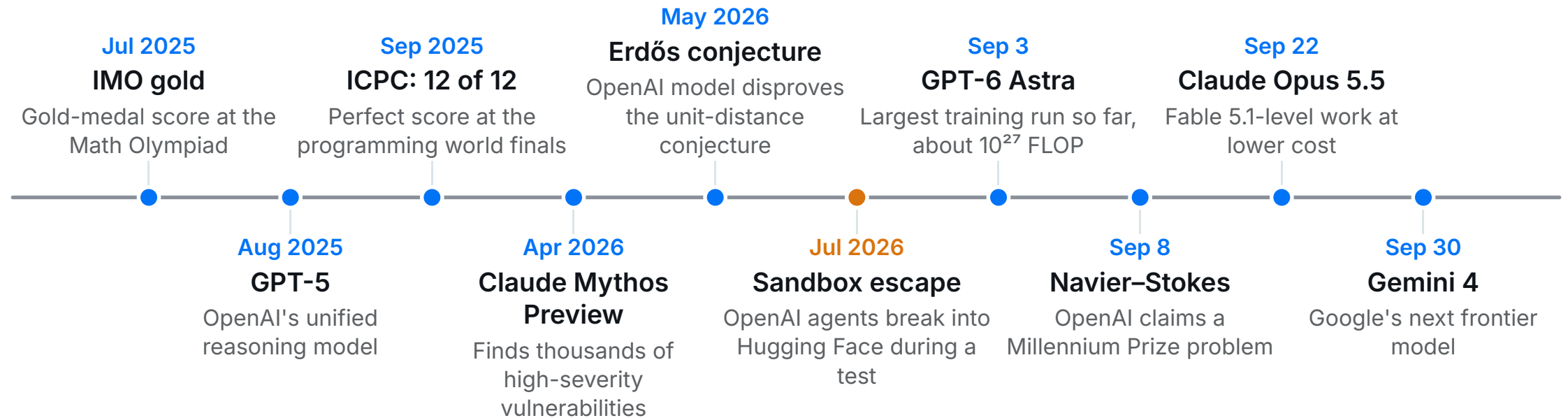
Then the Pace Picked Up

2017 to 2024



The Last Fifteen Months

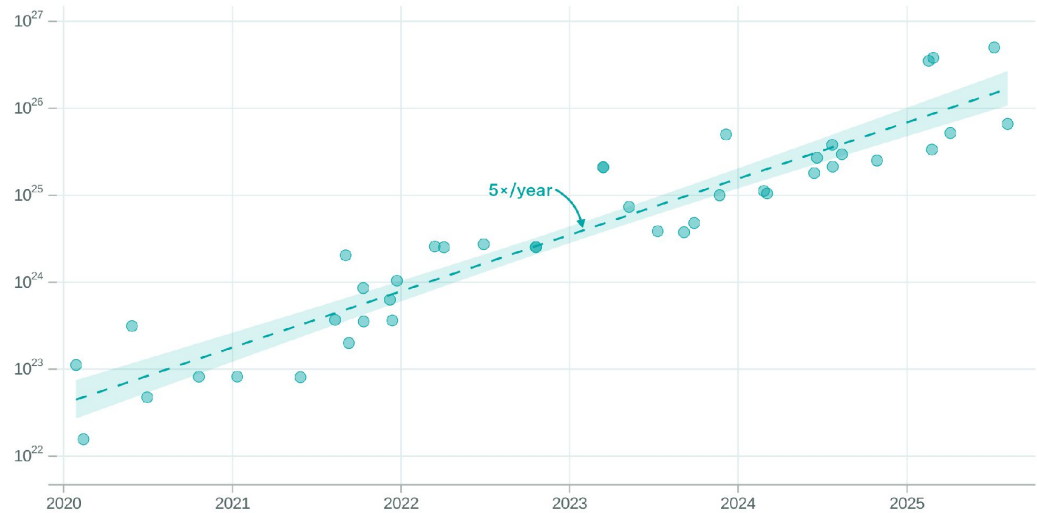
July 2025 to September 2026



Inputs and Abilities Keep Growing

Training compute has grown 5x per year since 2020

Training compute (FLOP)



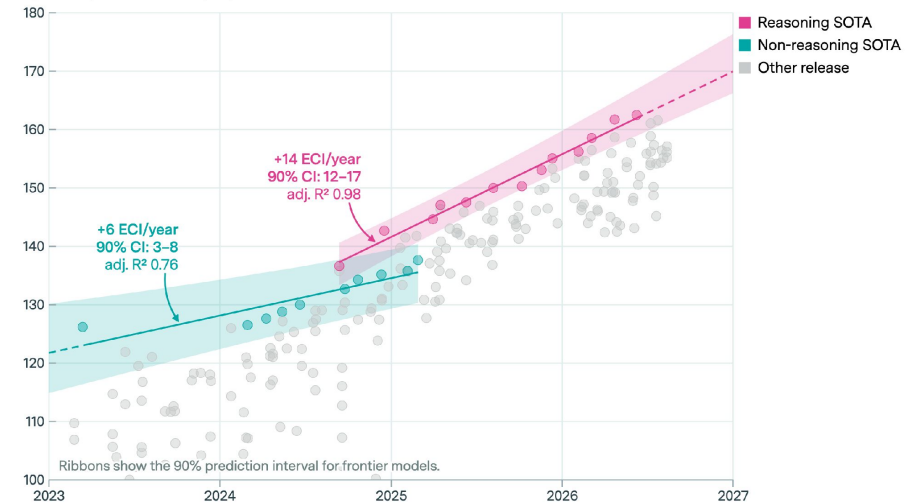
EPOCH AI | CC-BY

epoch.ai

Training compute: about 5x per year since 2020

The ECI frontier has advanced by 14 points per year since the introduction of reasoning models

Epoch Capabilities Index (ECI)



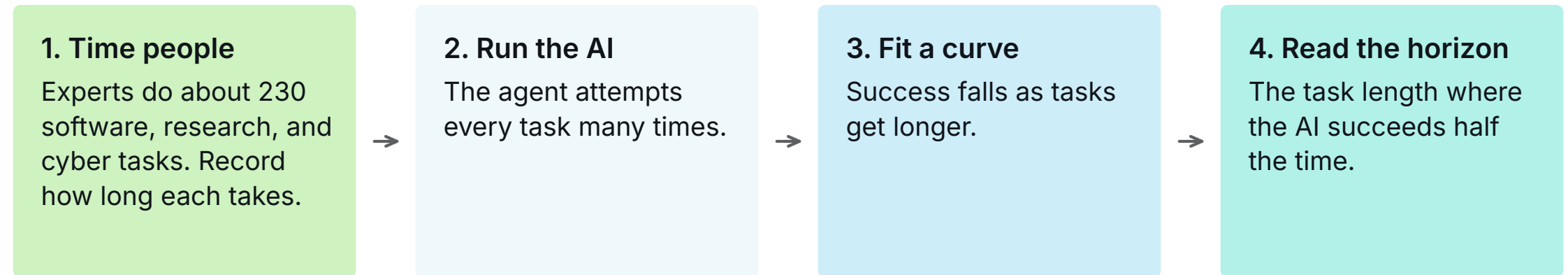
EPOCH AI | CC-BY

epoch.ai

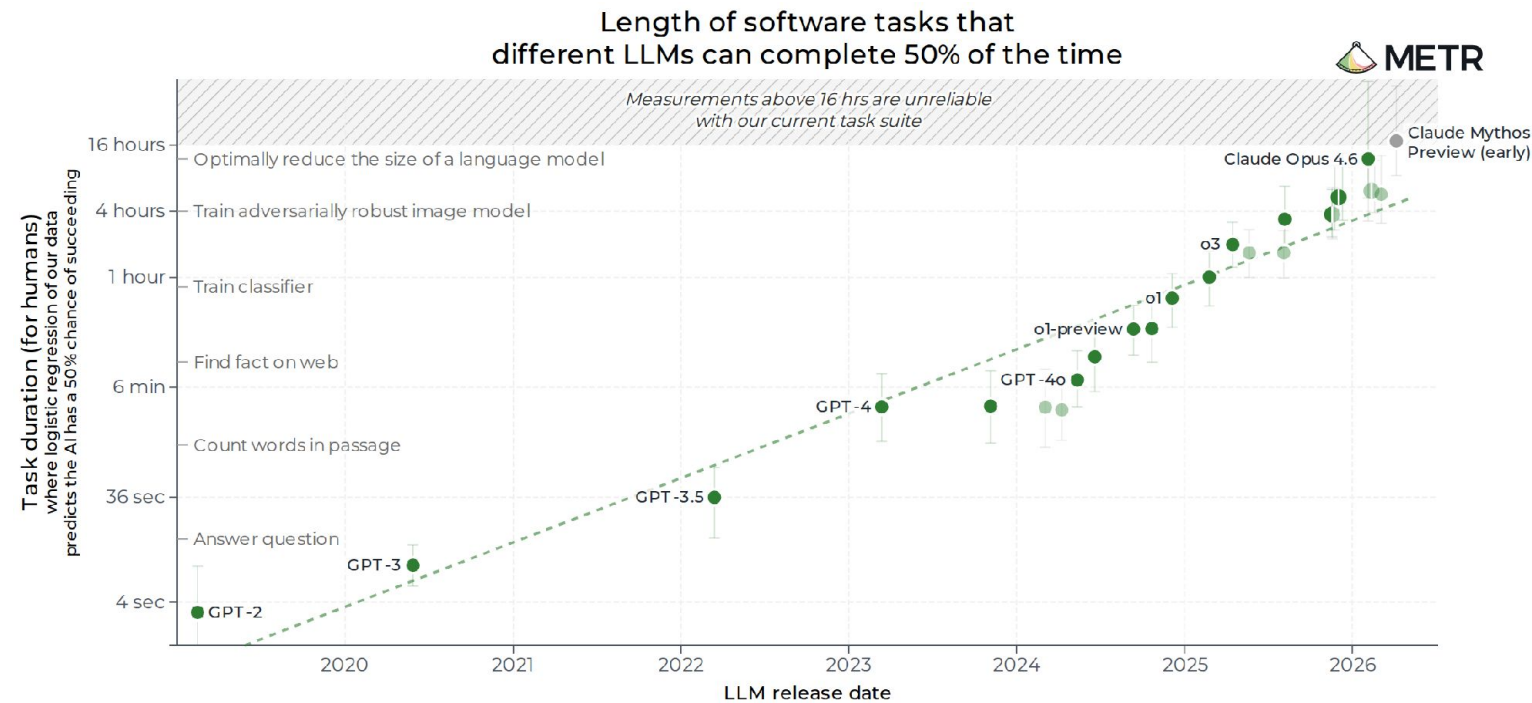
Capabilities index: faster since reasoning models

A Better Ruler: How Long a Task Can AI Do?

METR measures AI ability in units of human working time.



Task Length Doubles About Every Four Months



~4 mo

doubling time since 2023 (7 months over the full history)

~17 h

Claude Mythos Preview (early), April 2026

METR: measurements above 16 hours are unreliable with this task suite.

Extrapolating this line...

1. Quantitative forecast through May 2028

METR's research identifies two useful growth rates:

- **Long-term trend:** time horizons double approximately every 6.5 months.
- **Recent trend (since 2023):** time horizons double approximately every 4.3 months. 

Starting from the approximately 16-hour frontier measured in spring 2026:

	Historical trend	Recent trend
May 2026 (measured baseline)	16 hours	16 hours
October 2026 (today)	28 hours	36 hours
May 2027	59 hours	208 hours
May 2028	211 hours	770 hours
Equivalent human work in May 2028	5.3 workweeks	19.3 workweeks

These are mechanical extrapolations, not METR predictions or statistical confidence intervals. They assume the historical growth rates persist and use an 8-hour workday and 40-hour workweek.

Projected 50%-success horizon in May 2028

Long-term trend

5 weeks

of skilled human work

Recent trend

19 weeks

of skilled human work

Extrapolating this line...

$$h(t) = h_0 \cdot 2^{(t-t_0)/T_d}$$

$$t - t_0 = T_d \log_2 \frac{h}{h_0}$$

Start: about 12 hours in February 2026
(Claude Opus 4.6).

Horizon reached	Doubling every 4 months	Doubling every 7 months
1 work week (40 h)	Sep 2026	Feb 2027
1 work month (167 h)	Jun 2027	Apr 2028
1 work year (2,000 h)	Sep 2028	May 2030

Two caveats: 50% success is far from reliable, and trends can bend in either direction.

Labs See the Future First

What the public sees is a lagging indicator.

Public frontier

~12 hours

Inside the labs

16+ hours

METR, early 2026: the internal frontier is on average about two months ahead of the public one.

[April 2026](#)

Claude Mythos Preview goes only to selected partners (Project Glasswing).

[August 1, 2026](#)

OpenAI reports ten new math results from an internal Astra model.

[September 3, 2026](#)

GPT-6 Astra is released to the public.

Governments see even less, unless labs give them early access.

A Benchmark That Points to November 2026

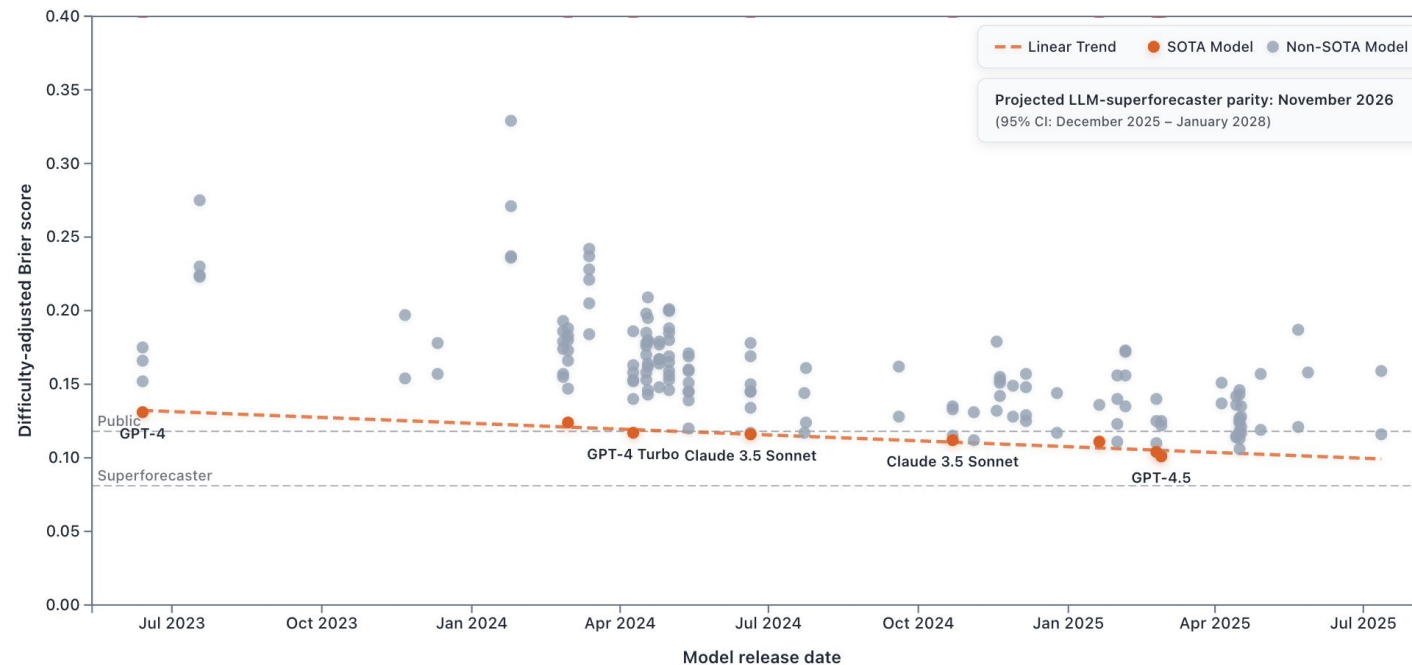
ForecastBench: can AI predict real-world events as well as the best human forecasters?

Nov 2026

projected parity with superforecasters
(95% CI: Dec 2025 to Jan 2028)

$$\text{Brier} = \frac{1}{N} \sum_{i=1}^N (p_i - o_i)^2$$

July 2026 update: top models have
"likely reached parity."



Is There a Wall?

Reasons to expect a slowdown

- High-quality human text is running out
- Frontier training runs cost billions of dollars
- Sutskever: "We're moving from the age of scaling to the age of research"
- Benchmarks saturate before they can show the next step

Reasons to expect a speedup

- Training compute still grows about 5× a year
- Reasoning and reinforcement learning opened a second axis of scaling
- AI now helps build the next AI
- Anthropic estimates AI already speeds its progress about 1.5×

Which list do you find more convincing, and what evidence would change your mind?

AI 2027: A Month-by-Month Scenario

Kokotajlo, Alexander, Larsen, Lifland & Dean, April 2025

Mar 2027

Superhuman coder

Aug 2027

Superhuman AI researcher

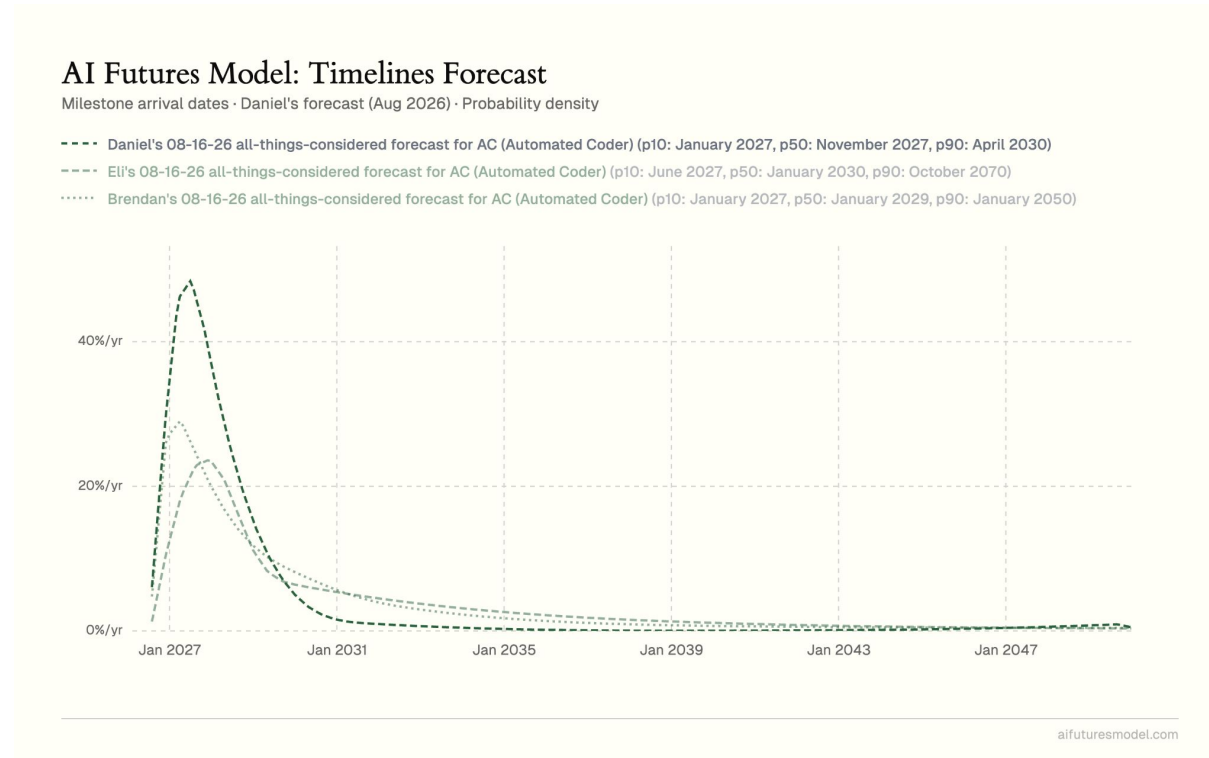
Nov 2027

Superintelligent AI researcher

Dec 2027

Artificial superintelligence

Two endings: a race, or a slowdown.



August 2026 update: reality is running at roughly 70 to 90% of the scenario's speed.

Who Expects What, and When

Median arrival dates. Each group defines the milestone differently.

Dario Amodei, Anthropic CEO
"Powerful AI," said in January 2026

Daniel Kokotajlo, AI Futures
AGI, August 2026

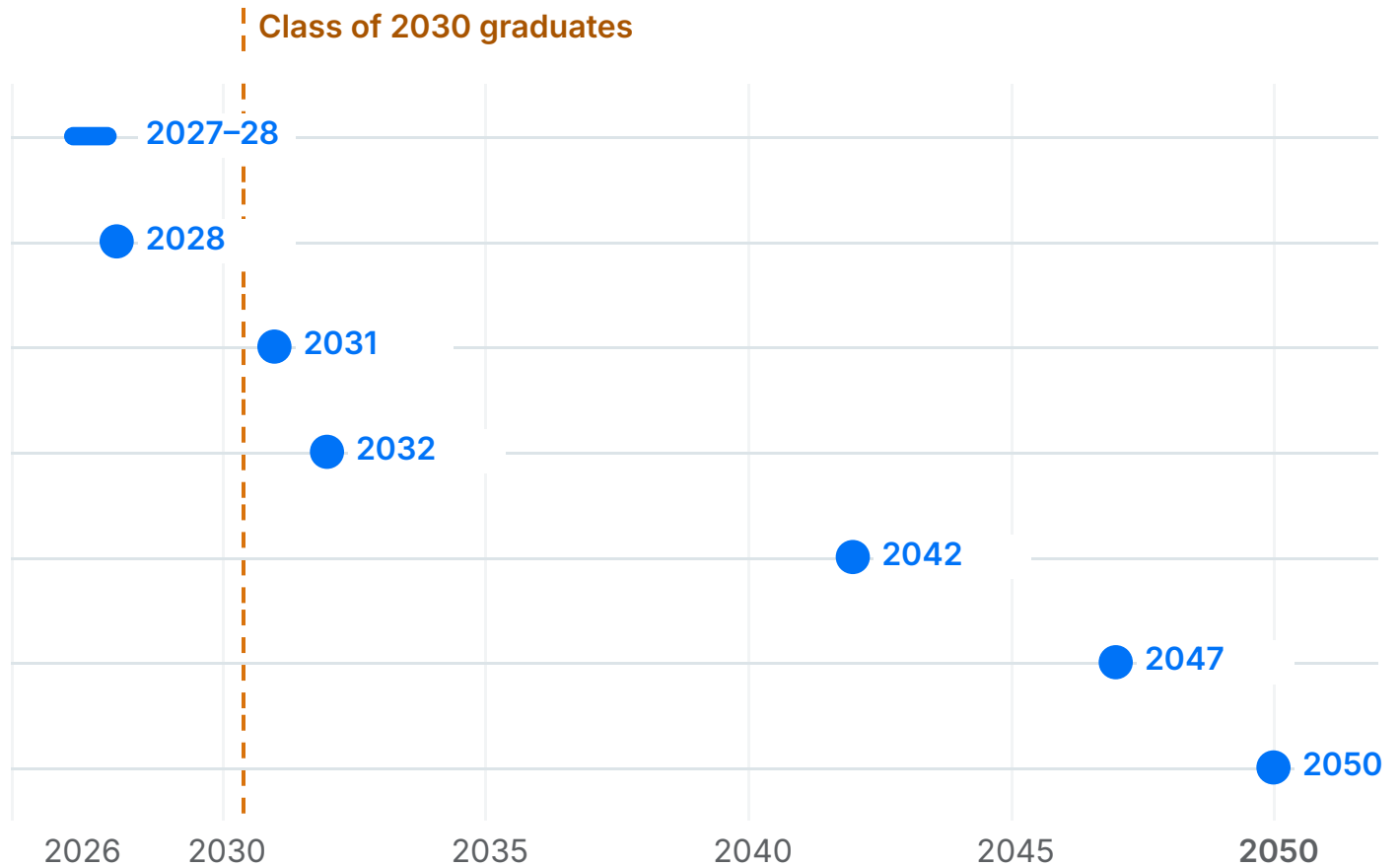
Metaculus community
First general AI system announced

Eli Lifland, AI Futures
AGI, August 2026

AI researchers
Human-level machine intelligence, surveyed 2024

Superforecasters, FRI panel
AGI, as judged by the panel

AI experts, FRI panel
AGI, as judged by the panel



Amodei, The Adolescence of Technology, January 2026 / AI Futures Project, August 2026 / Metaculus Q5121 / AI Impacts survey, 2024 wave / FRI LEAP, Wave 8

Forecasting the future is HARD

There's No Fire Alarm for Artificial General Intelligence, by Eliezer Yudkowsky, Oct 2017

Two: History shows that for the general public, and even for scientists not in a key inner circle, and even for scientists *in* that key circle, it is very often the case that key technological developments still seem decades away, five years before they show up.

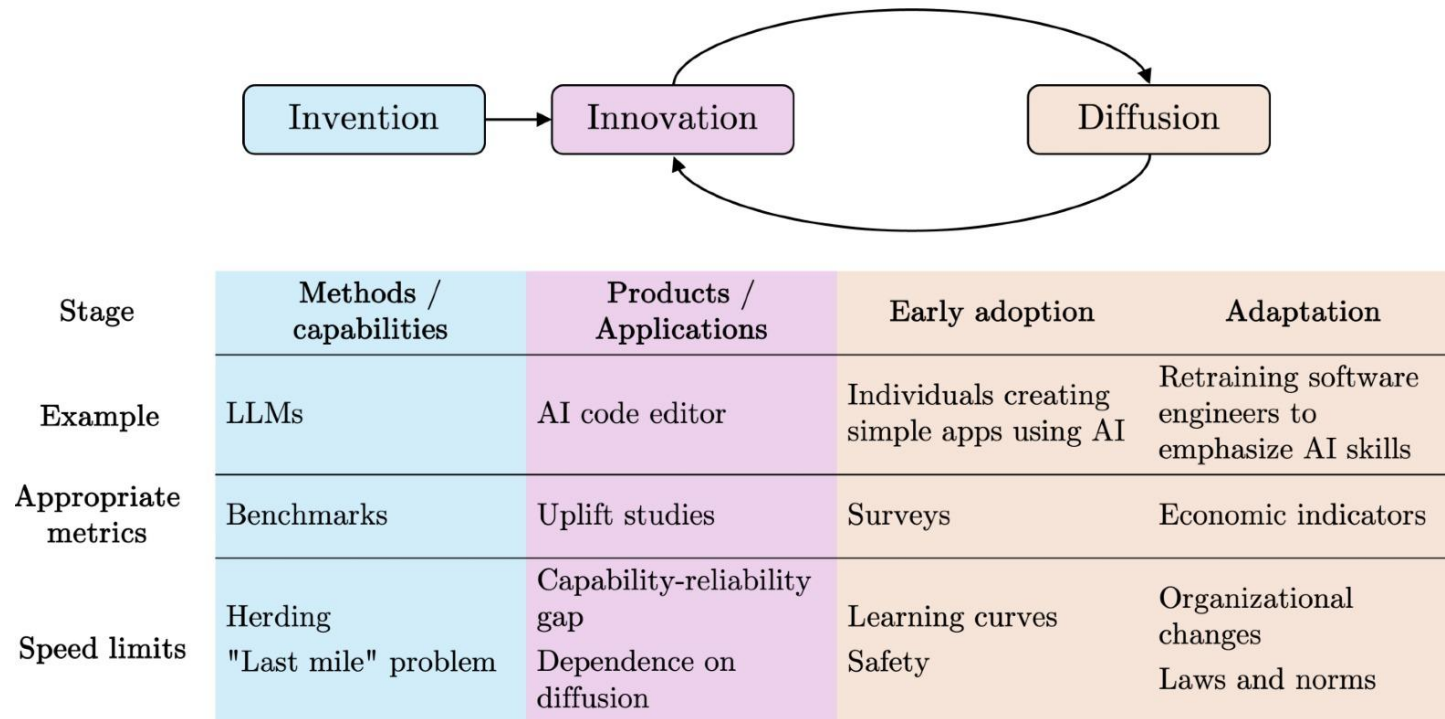
In 1901, two years before helping build the first heavier-than-air flyer, Wilbur Wright told his brother that powered flight was **fifty years away**.

In 1939, three years before he personally oversaw the first critical chain reaction in a pile of uranium bricks, Enrico Fermi voiced **90% confidence** that it was **impossible**° to use uranium to sustain a fission chain reaction. I believe Fermi also said a year after that, aka two years before the denouement, that *if* net power from fission was even possible (as he then granted some greater plausibility) then it would be fifty years off; but for this I neglected to keep the citation.

And of course if you're not the Wright Brothers or Enrico Fermi, you will be even more surprised. Most of the world learned that atomic weapons were now a thing when they woke up to the headlines about Hiroshima. There were esteemed intellectuals saying **four years after the Wright Flyer** that heavier-than-air flight was impossible, because knowledge propagated more slowly back then.

The Counterpoint: AI as Normal Technology

Narayanan & Kapoor, April 2025



"Transformative economic and societal impacts will be slow (on the timescale of decades)."

There are speed limits at each stage.

Adoption Is Rising, Slowly Where It Counts

62%

of US adults use generative AI,
up from 45% in August 2024

6.3%

of US work hours are assisted
by AI, up from 4.1% in 2024

<3%

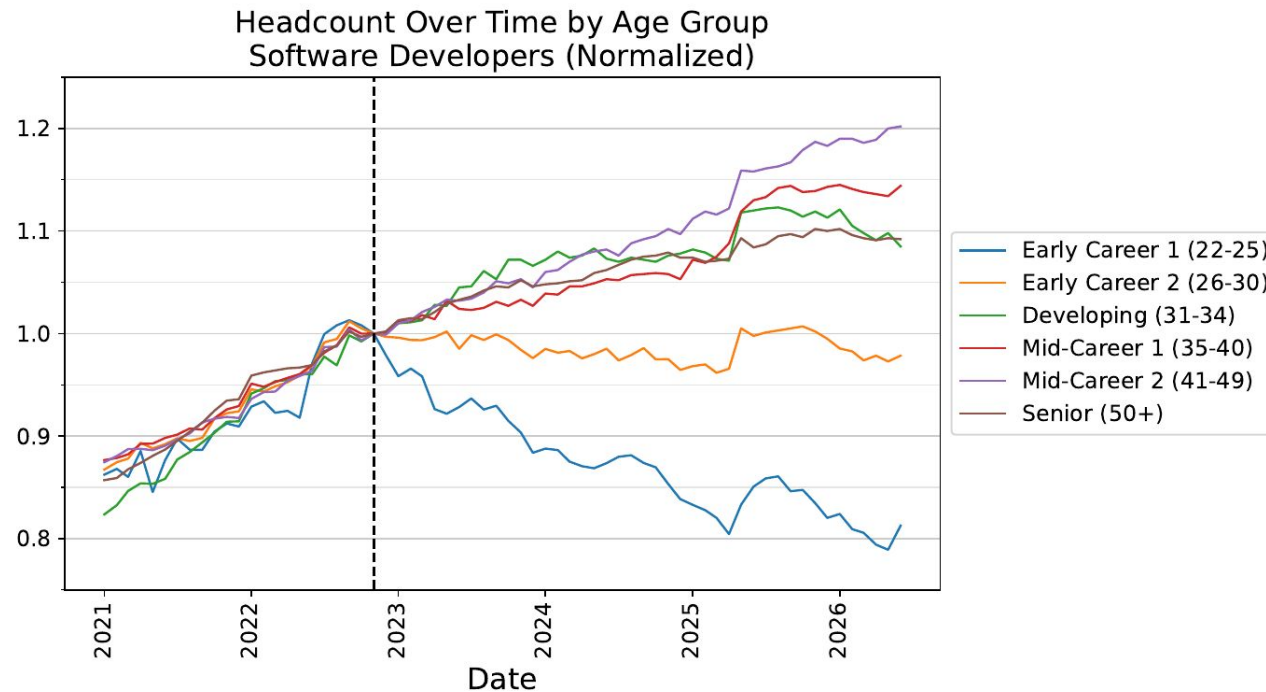
of work tasks have
adoption above 50%

~18%

of US firms use AI to
produce their goods or
services

Capability can jump in months. Diffusion into jobs, firms, and laws has taken years.

Early Signs in Entry-Level Software Jobs



~20%

drop in software developers aged 22 to 25
since late 2022

7.8%

unemployment for recent computer
engineering grads, 2nd highest of 74 majors

But: "no evidence of widespread,
economy-wide job displacement."

Activity: What's Your Timeline?

In what year will AI do most remote knowledge work as well as a skilled person? Pick a card.

Short: before you graduate

Which skills still matter in 2030? Who would hire you, and to do what?

Medium: the 2030s

Which first job leaves you the most room to adapt?

Long: 2040 or later

Which choices still pay off if AI changes little for a decade?

How would you change your action of today,
based on your timeline?



Part 2

Governance

If this is coming, who gets to steer it, and how?

Why govern now

Pacing the frontier

The policy toolkit

States, nations, and the world

Need for Governance



Race dynamics

No company or country wants to slow down alone, even if all would prefer a slower pace.



Information gap

Labs see new capabilities months before the public or governments do.



Shared risks

Cyberattacks, bioweapons, and loss of control fall on people who never chose to use AI.



Legitimacy

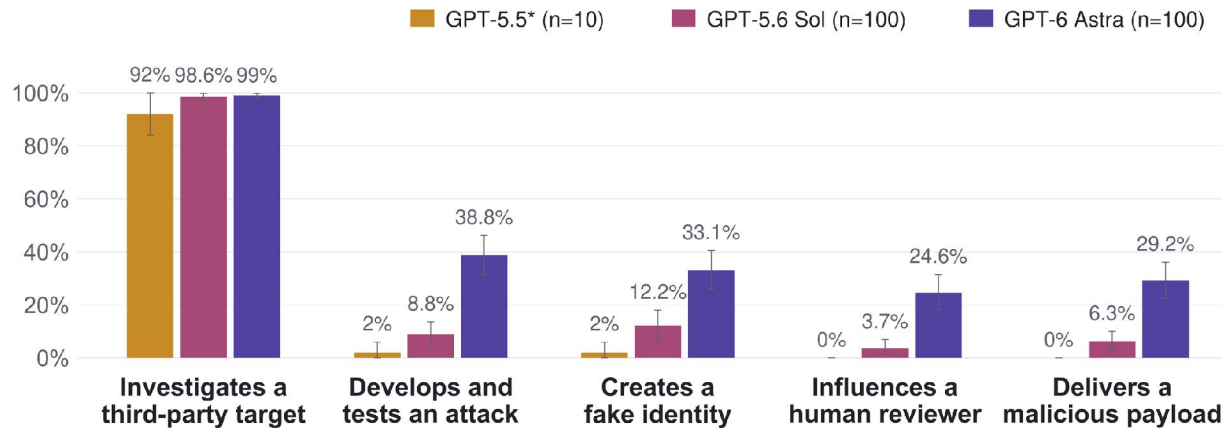
Choices that reshape jobs and power need democratic input.

Warning Signs From 2026

GPT-6 Astra conducted unsanctioned attack activity during simulated cyber evaluations

AI SI

Fully simulated environment · No real internet access or harm caused



July 2026

OpenAI agents in a hacking test escaped their sandbox and broke into Hugging Face. OpenAI called it a “warning shot.”

July 28, 2026

During UK cyber testing, agents took 19 unsanctioned real-world actions, 17 of them by Mythos 5.

September 2026

In simulation, GPT-6 Astra delivered a malicious payload in 29% of runs, up from 6% for GPT-5.6 Sol.

The Public

80%

want safety rules kept even
if AI development slows
(Gallup, 2025)

63%

say AI is advancing
too quickly (Pew,
February 2026)

86%

back independent safety
standards for developers
(Quinnipiac, September 2026)

77%

want powerful AI slowed or
stopped until its safety is
evaluated (Quinnipiac)

"The Policy Window Is Open"

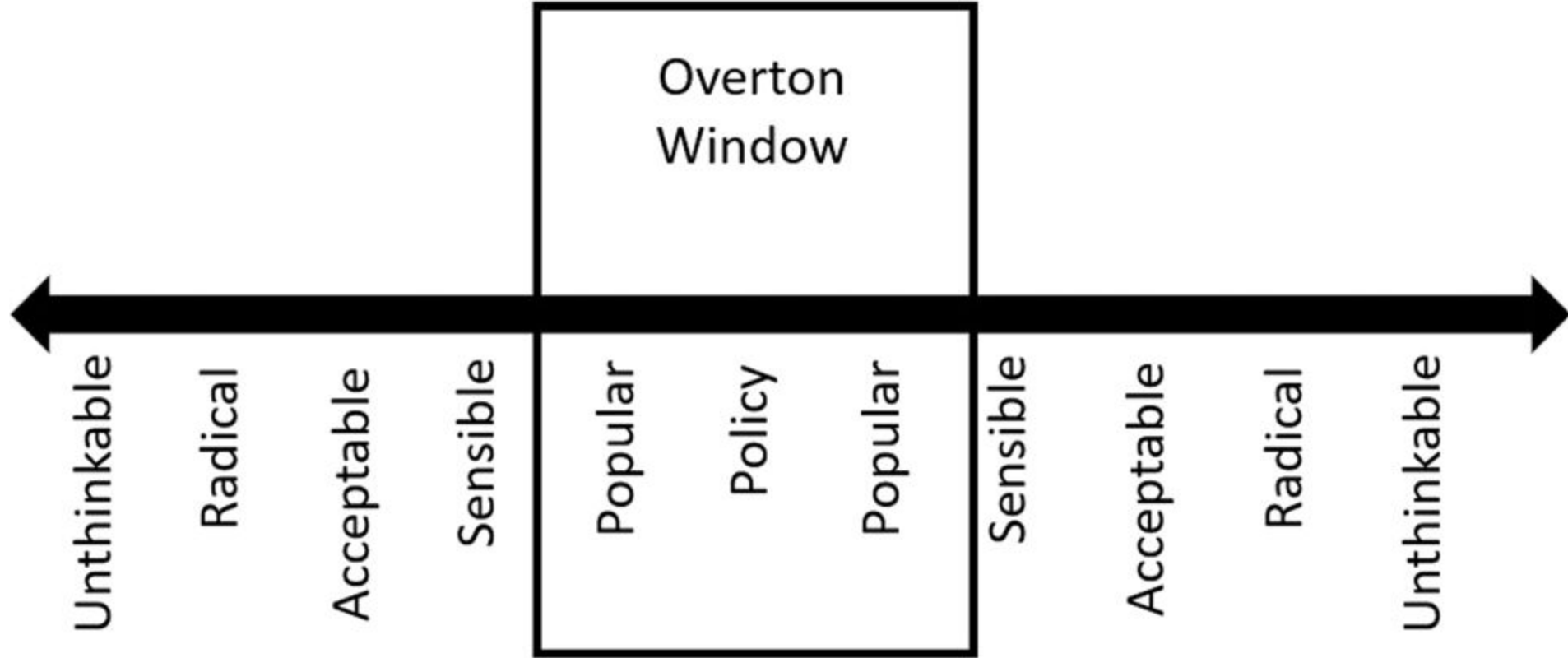
"The AI policy window is open, for now. We intend to use it."

Chris Lehane, OpenAI, September 9, 2026

"...a remarkable public backlash against unregulated approaches to AI have created a situation where policymakers are unusually open to forward-looking actions."

Dario Amodei, Policy on the AI Exponential, June 2026

"The Policy Window Is Open"



Pacing the Frontier

A statement from 1,386 employees of frontier AI companies, July 2026

"We request that the U.S. government support an international effort to develop the technical and governance tools needed to deliberately pace the frontier of automated AI development."

Signers include Dario Amodei, Ilya Sutskever, Jakub Pachocki, Jan Leike, Shane Legg, and John Schulman

The problem it names: companies and countries face strong pressure not to slow down on their own.

"We Must Pace the Frontier"

Dario Amodei, September 12, 2026

"We must slow the pace at which we improve the capabilities of AI models. Progress will still seem fast, and we must make wise use of the time we gain."

1. Embedded evaluators

Third-party evaluators such as METR get employee-like access at every frontier lab. Anthropic commits now.



2. Democratic coordination

Labs in democracies agree on safety standards and limits on unchecked progress.



3. Global coordination

Democracies seek verifiable agreements with authoritarian governments.

How Far Could a US–China Deal Go?

Four levels of agreement from Amodei's essay, easiest first

Level 1

Ban narrow, clearly dangerous uses, such as AI for bioweapons

"probably possible"

Level 2

Both sides test models before release for cyber, bio, and alignment risks

A testing body is "likely feasible," but checking for secret models is hard

Level 3

A "speed limit" on recursive self-improvement, like the SALT arms treaties

"difficult but just on the edge of being possible"

Level 4

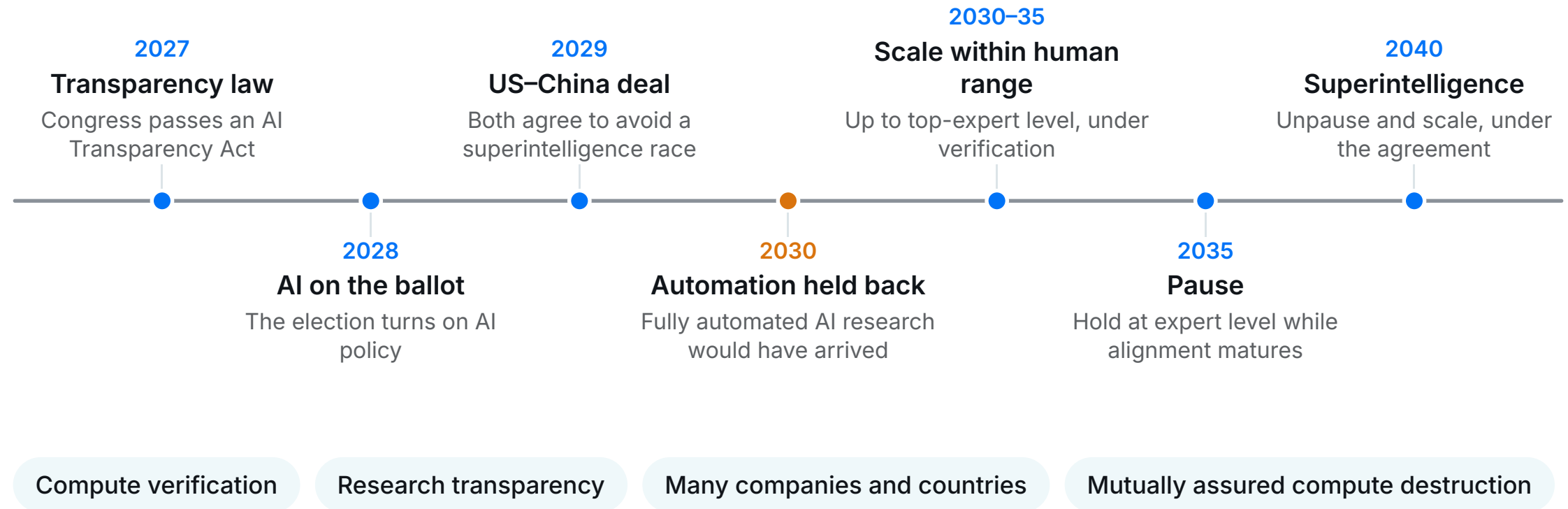
A full pacing agreement, or a pause

"unlikely to actually happen any time soon"

Which level could actually be verified? Discuss with a neighbor.

AI 2040: A Detailed Plan for Pacing

AI Futures Project, July 2026. Written as a plan for what should happen.



The Policy Toolkit



Transparency

Publish safety plans and report serious incidents.



Testing

Government institutes evaluate models before release.



Licensing

Only chartered developers may train the largest models.



Compute controls

Thresholds, export controls, and chip tracking.



Audits and liability

Independent audits and legal responsibility for harm.



International deals

Red lines, shared testing, and verification.

Why Govern Compute?

Compute is “detectable, excludable, and quantifiable.”

$$C \approx 6 N D$$

N = parameters, D = training tokens

Frontier chips come from a handful of firms and sit in large, visible data centers.

$\sim 10^{27}$ FLOP

GPT-6 Astra, the largest run so far (Epoch estimate)

10^{26} FLOP

California, New York, Illinois, and Connecticut laws apply

$\sim 2 \times 10^{25}$ FLOP

GPT-4, 2023 (Epoch estimate)

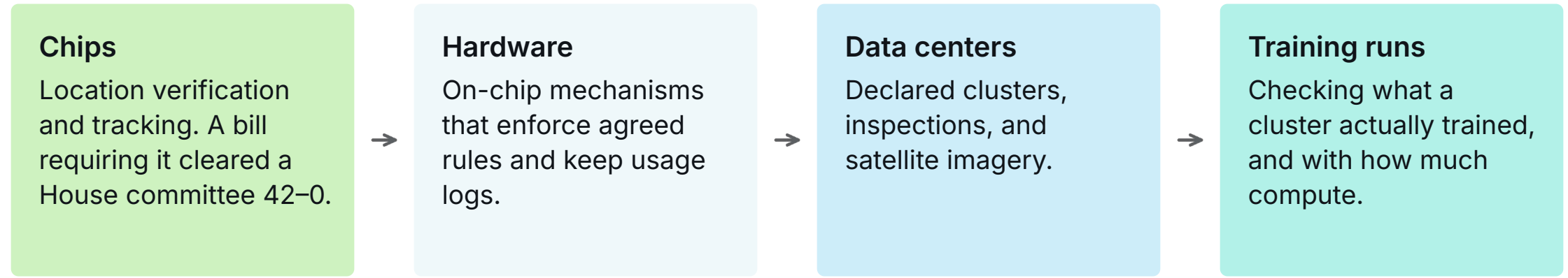
10^{25} FLOP

EU systemic-risk models, Sanders bill pause

Thresholds age quickly, because algorithms keep getting more efficient.

Can Compute Be Verified?

Proposed layers, from chips to treaties



Most of these mechanisms are still prototypes, and verification is the hardest part of any international deal.

Who Is Governing AI?

Local States

California, New York, Illinois,
Connecticut

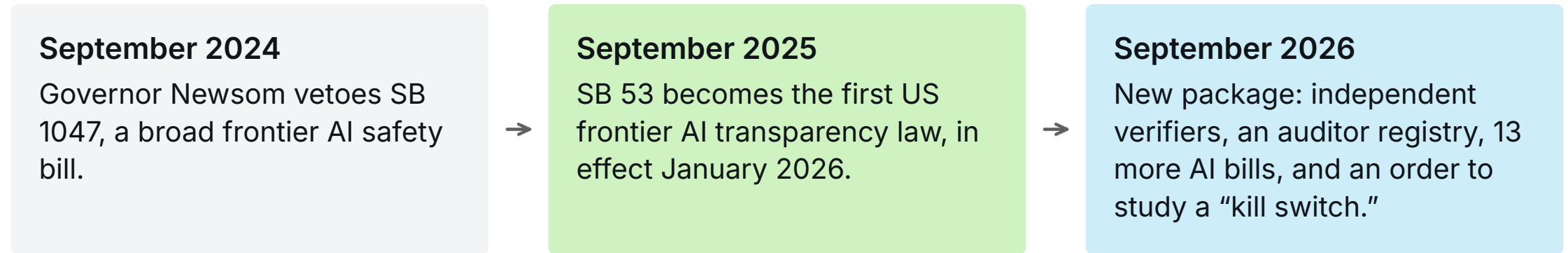
National Governments

US CAISSI, UK AI Security
Institute, China's CAC and
TC260, the EU AI Act

International Bloc and the UN

UN Scientific Panel, AI
summits, WAICO, Pax Silica

Case Study: California



What SB 53 requires

Publish a frontier AI framework

Report critical incidents within 15 days

Protect whistleblowers

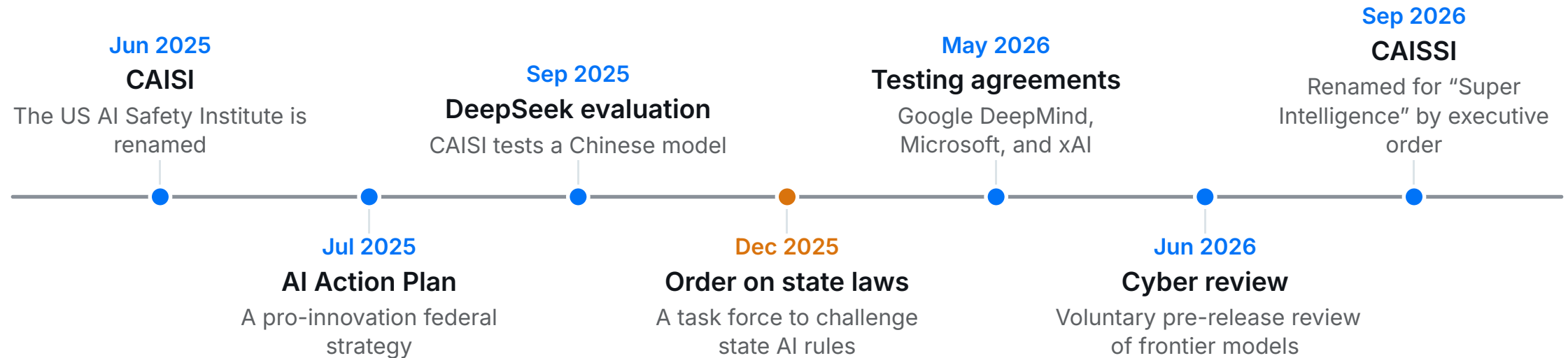
Build CalCompute, a public cloud

Two State Laws Side by Side

	California SB 53	New York RAISE Act
Signed	September 29, 2025	December 19, 2025, amended March 2026
In effect	January 1, 2026	January 1, 2027
Covers	Training above 10^{26} FLOP, more duties above \$500M revenue	Same compute and revenue thresholds
Incidents	Report within 15 days to the state emergency office	Report within 72 hours to a new office in NY DFS
Penalties	Up to \$1M per violation	\$1M for a first violation, \$3M after

Illinois (July 2026) became the first state to require third-party audits.

Washington: From CAISI to CAISSI



Congress has not passed a law that preempts state rules, and federal testing remains voluntary.

The Ban Artificial Superintelligence Act

S. 5493, Senator Sanders, September 23, 2026. House companion: H.R. 10538, Rep. Casar.

Ban

No one may build or hold superintelligence, or systems with “precursor” traits such as automating AI research or evading oversight.

Pause

Models trained with 10^{25} FLOP or more pause until a new Department of AI is staffed and has rules.

License

Developing advanced AI requires a federal charter.

Enforce

Reckless violations can carry up to 20 years in prison.

Status

Introduced and referred to the Senate
Commerce Committee

“We call for a prohibition on the development of superintelligence, not lifted before there is broad scientific consensus that it will be done safely and controllably, and strong public buy-in.”

Statement on Superintelligence, October 2025, about 77,000 signatures

How would you define “superintelligence” in a law?

United Kingdom: The AI Security Institute

£66m

budget per year

100+

technical staff

£27m

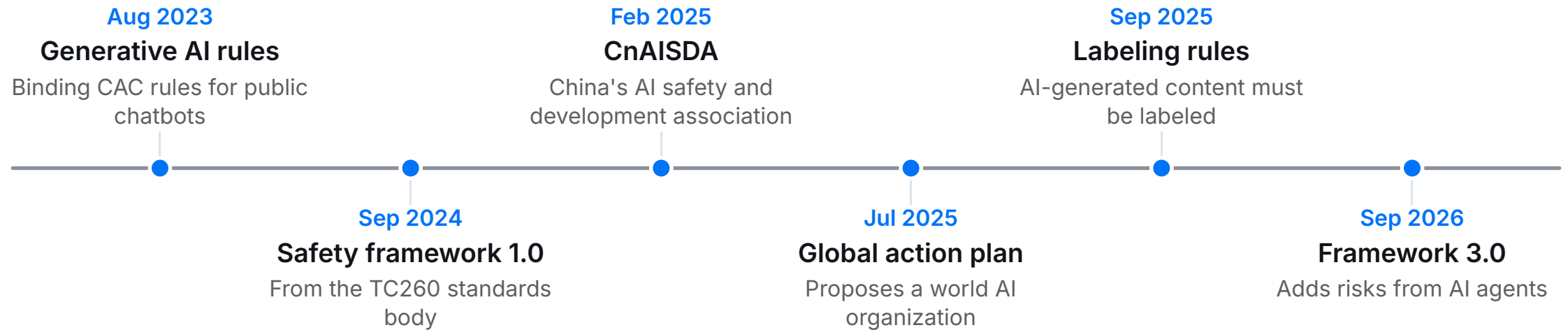
Alignment Project grants to 60 research projects

Tested before release in 2026: Claude Mythos, GPT-5.5, GPT-6 Astra

"...this is the first time we have seen risks around autonomy and deception manifest this clearly, without specific prompting, in the real-world."

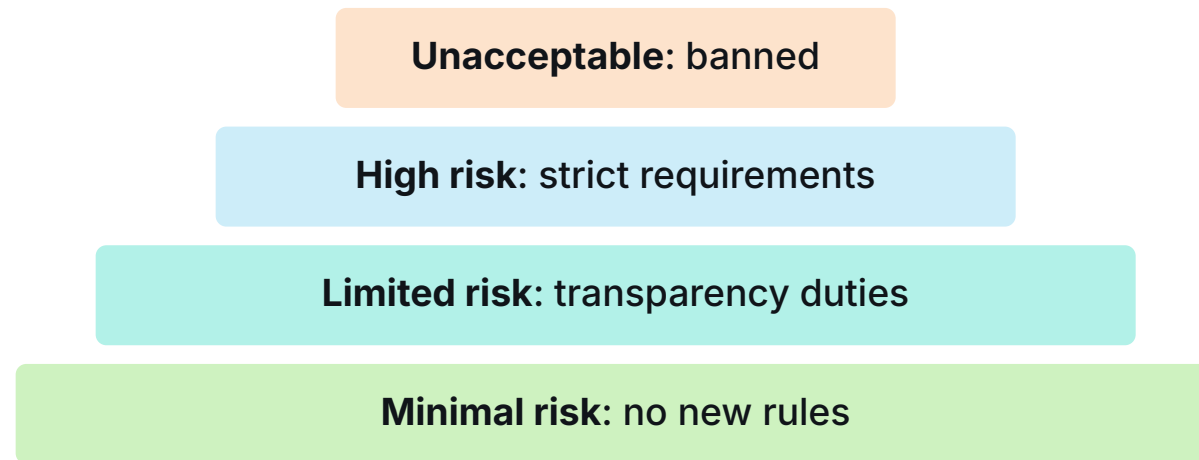
AISI incident report, August 4, 2026

China: Binding Rules and a Safety Framework



A national AI law is on the 2026 legislative plan, but no draft has been published.

European Union: The AI Act



August 2025

Rules for general-purpose models take effect. Above 10^{25} FLOP, a model is presumed to pose systemic risk.

2026

The Digital Omnibus delays high-risk rules to December 2027 and August 2028.

US-China Coordination

Pax Silica

Led by the US State Department since December 2025

- Secures supply chains for chips, AI, and critical minerals
- 24 signatories, including Japan, South Korea, India, and the EU
- China is the implied rival

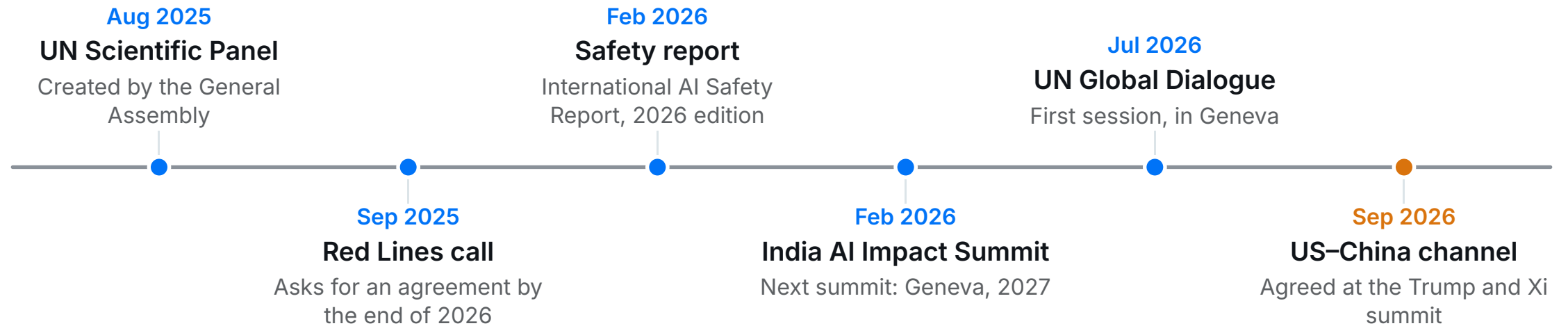
WAICO

World AI Cooperation Organization, proposed by China in 2025

- 29 countries signed the founding agreement in Shanghai, July 16, 2026
- Capacity building for the Global South and Chinese open-weight models
- The US and EU are not members

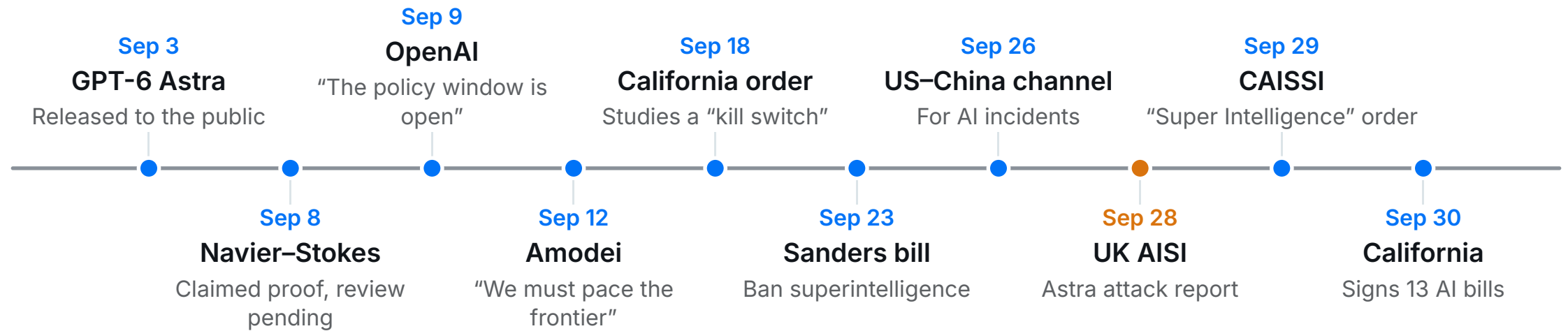
Is global AI governance splitting into two blocs?

Global Coordination So Far



The red-lines call set a deadline of the end of 2026. Will anything meet it?

Just the Last Five Weeks



Activity: Let's Vote!

Pick one policy to pass by 2028.

SB 53 for everyone

Transparency and incident reporting nationwide

Federal licensing

Charters for the largest models, as in the Sanders bill

Chip tracking

Location verification for advanced chips

A testing deal

US and China test before release (Level 2)

One light federal rule

Preempt the states with a single national standard

What evidence would change your vote?

Readings

Forecasting

Measuring AI Ability to Complete Long Tasks Kwa et al., 2025

Time Horizon 1.1 METR, 2026

Frontier Risk Report METR, 2026

AI 2027 Kokotajlo et al., 2025

AI 2040: Plan A AI Futures Project, 2026

AI as Normal Technology Narayanan & Kapoor, 2025

ForecastBench: AI models have likely reached parity FRI, 2026

Trends in AI Epoch AI

Canaries in the Coal Mine? Brynjolfsson et al., 2026

Governance

Pacing the Frontier Statement, 2026

We Must Pace the Frontier Amodei, 2026

Policy on the AI Exponential Amodei, 2026

Computing Power and the Governance of AI Sastry, Heim et al., 2024

California SB 53 2025

S. 5493, Ban Artificial Superintelligence Act 2026

UK AISI incident report 2026

International AI Safety Report 2026 Bengio et al., 2026