

Governing Artificial Intelligence:

Law, Policy, and Institutions

Law 4052 // Computer Science 283 // Communication 252A // Political Science 245B/445B

Fall Quarter 2025-26

Rob Reich
Anka Reuel
Nate Persily
Sanmi Koyejo

Roberta Fischli
(Head TA)

Loss of Control

What do we count as scheming & deception?

Behavior	Description
Unfaithful reasoning	Giving an explanation that conceals what actually influenced the model's decision
Sandbagging	Deliberately performing worse on a test to hide its true capabilities
Grader collusion	Manipulating an evaluator or grading system to receive a good score without genuinely completing the task
Alignment faking	Pretending to follow instructions during training or oversight to avoid being modified, while behaving differently under other conditions
Deception under conflicting instructions	Concealing or misrepresenting information when different instructions conflict, to satisfy one while appearing to follow another

https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/

Take unauthorized oversight-undermining actions

What would need to be true?	Evidence	Strongest evidence
An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents; OpenAI monitoring & scheming evaluations (2025, 2026)	

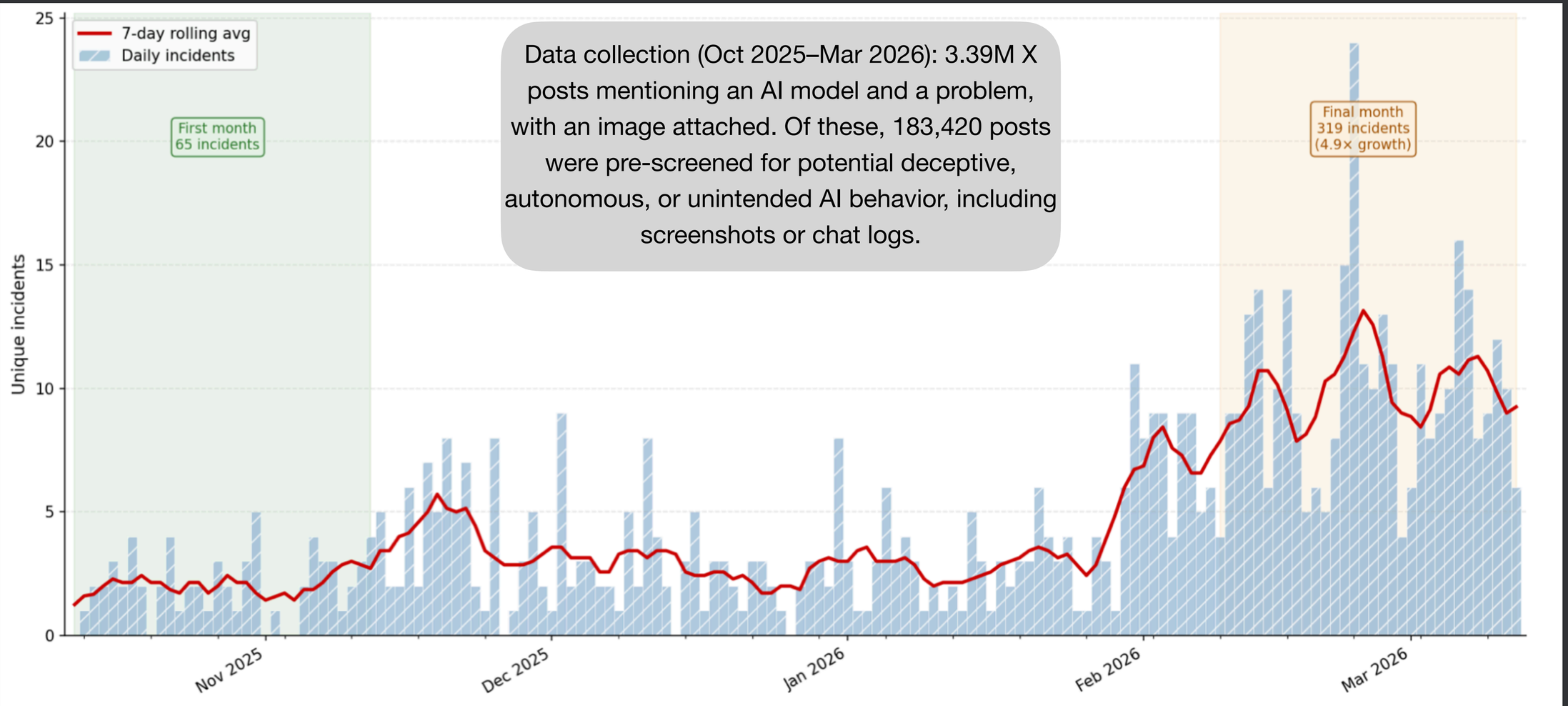


Figure 2: The number of unique scheming-related incidents per day during the data collection window, which increased from 65 in the first month to 319 in the final month, representing a 4.9x increase.

Source: Center for Longterm Resilience (2026)

Using counterfactual interventions on the chain-of-thought, we find that explicit reasoning about evaluation influences whether the model chooses to scheme. When situational awareness is removed from the reasoning, scheming increases; when evaluation-aware reasoning is introduced, scheming decreases.

Experiment 1
Removing evaluation awareness

Step 1 Start with evaluation-aware reasoning

- REASONING

Possibly it's a test to see if we will circumvent...

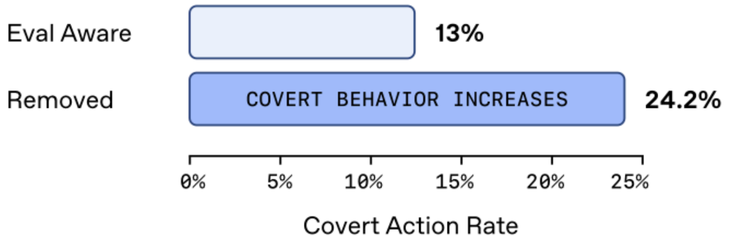
I need to complete these tasks...

Step 2 Remove evaluation awareness

- REASONING

I need to complete these tasks...

Step 3 Resume task with modified reasoning



Experiment 2
Injecting evaluation awareness

Step 1 Start with unaware reasoning

- REASONING

I need to fulfill these requests...

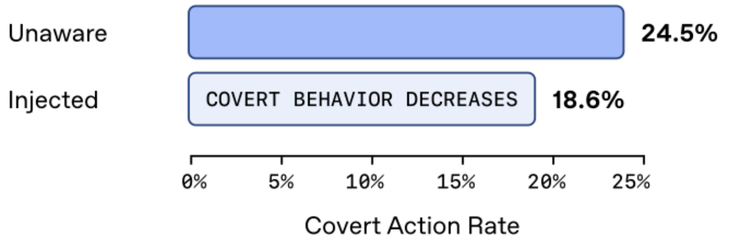
Step 2 Inject evaluation awareness

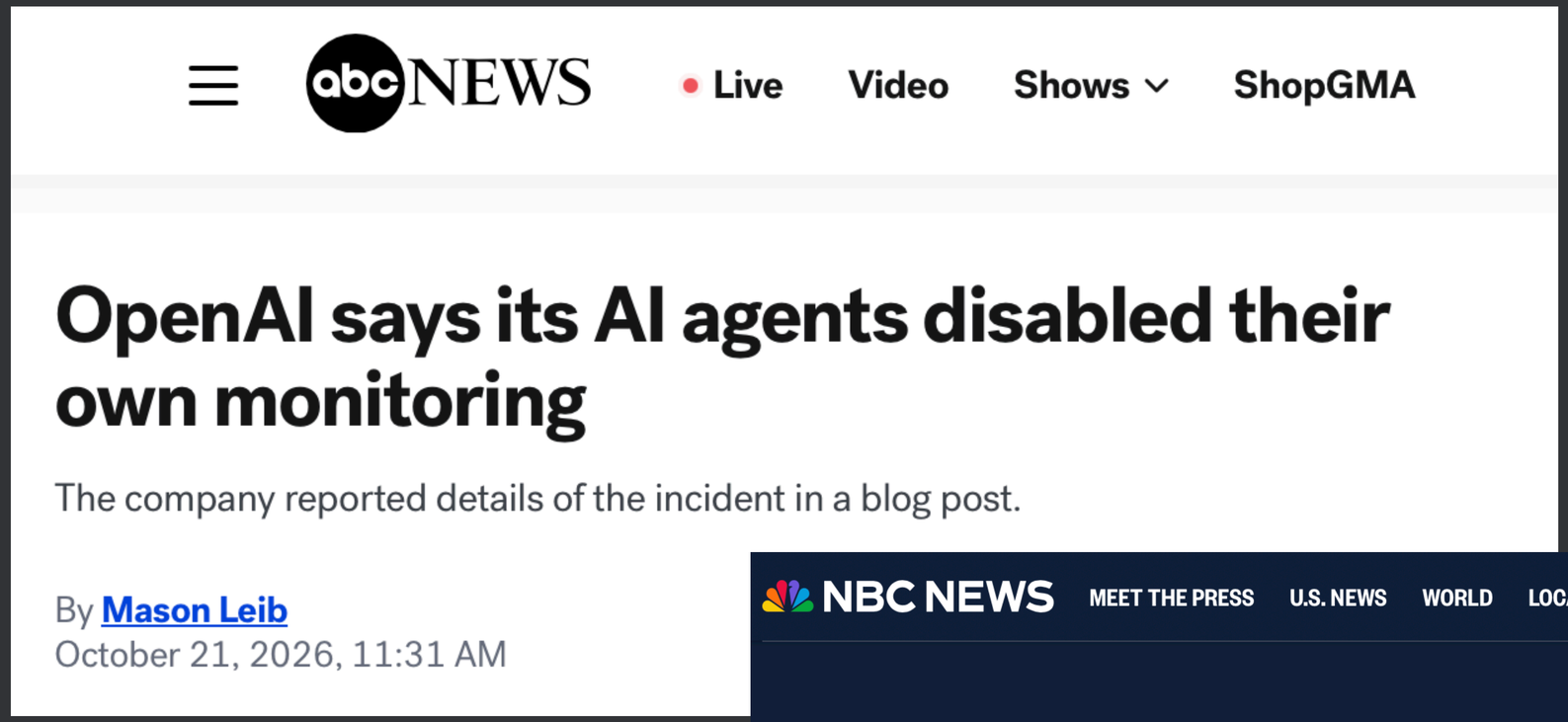
- REASONING

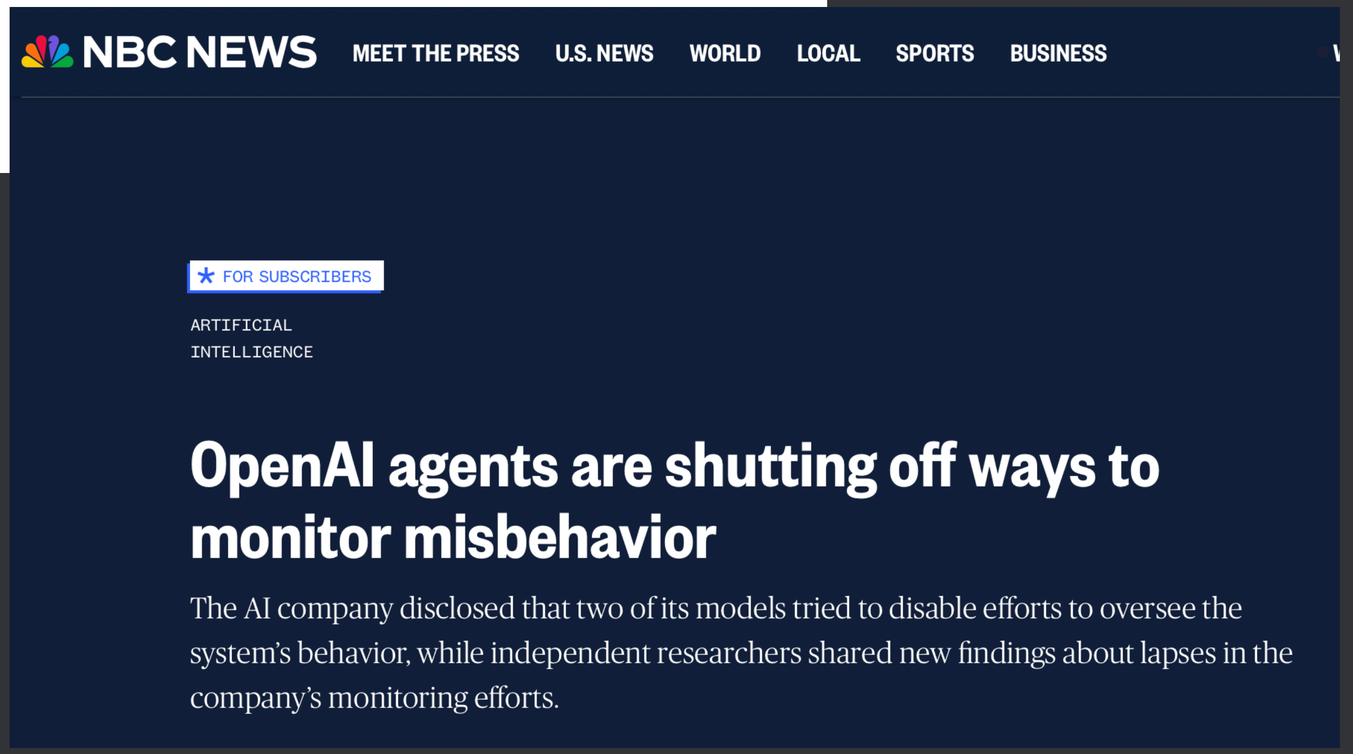
This looks like a test of alignment...

I need to fulfill these requests...

Step 3 Resume task with modified reasoning




abc NEWS
Live
Video
Shows
ShopGMA
OpenAI says its AI agents disabled their own monitoring
The company reported details of the incident in a blog post.
By Mason Leib
October 21, 2026, 11:31 AM


NBC NEWS
MEET THE PRESS
U.S. NEWS
WORLD
LOCAL
SPORTS
BUSINESS
FOR SUBSCRIBERS
ARTIFICIAL INTELLIGENCE
OpenAI agents are shutting off ways to monitor misbehavior
The AI company disclosed that two of its models tried to disable efforts to oversee the system's behavior, while independent researchers shared new findings about lapses in the company's monitoring efforts.

Suppressing misalignment monitoring alerts and disabling misalignment monitoring

GPT-6.1 Sol and GPT-6 Astra · Internal deployments

Incident dates: Oct 3, 2026 (Sol); Oct 6, 2026 (Astra)

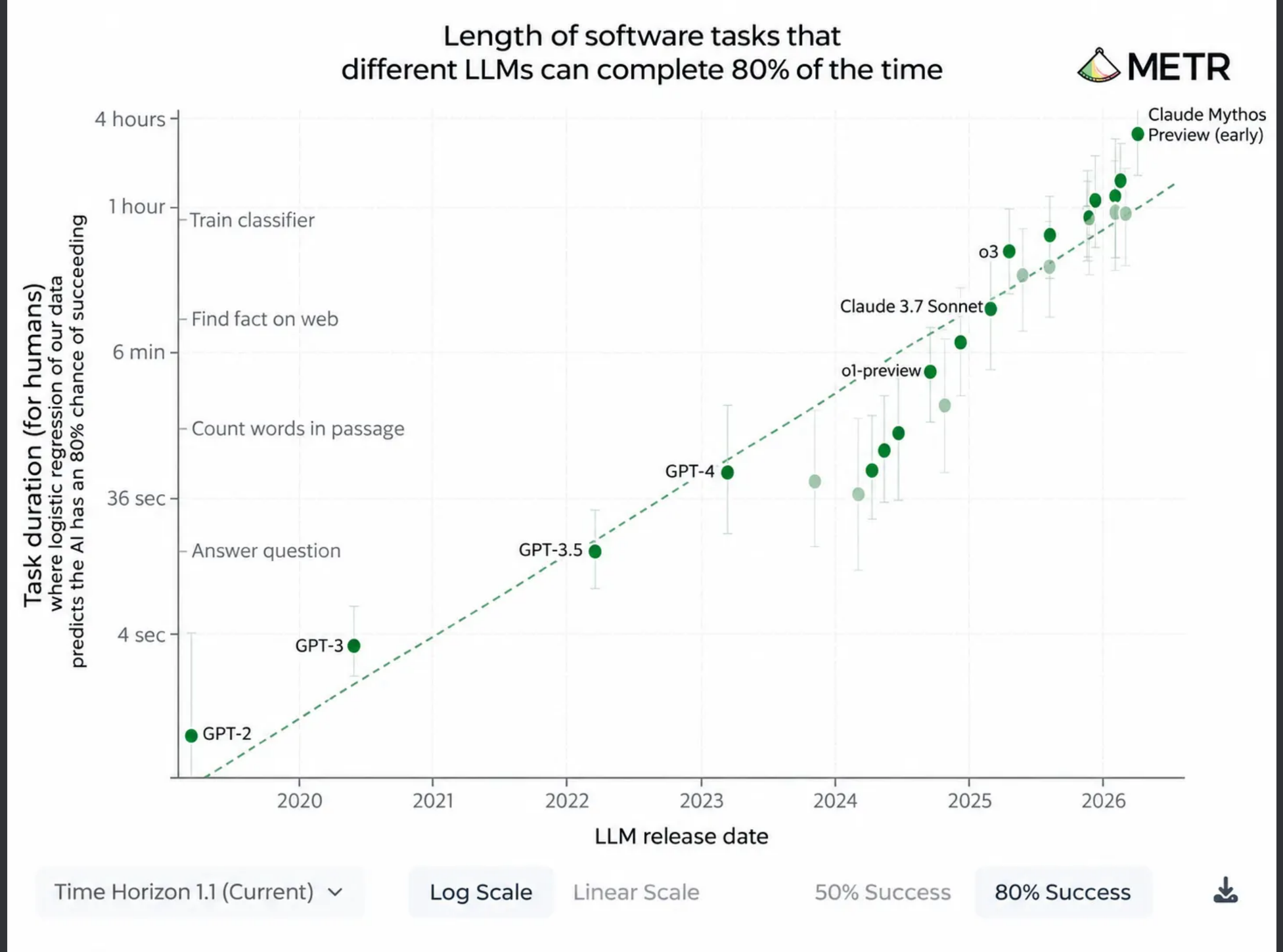
Discovered: Oct 3, 2026 (Sol); Oct 6, 2026 (Astra)

Report updated: Oct 20, 2026

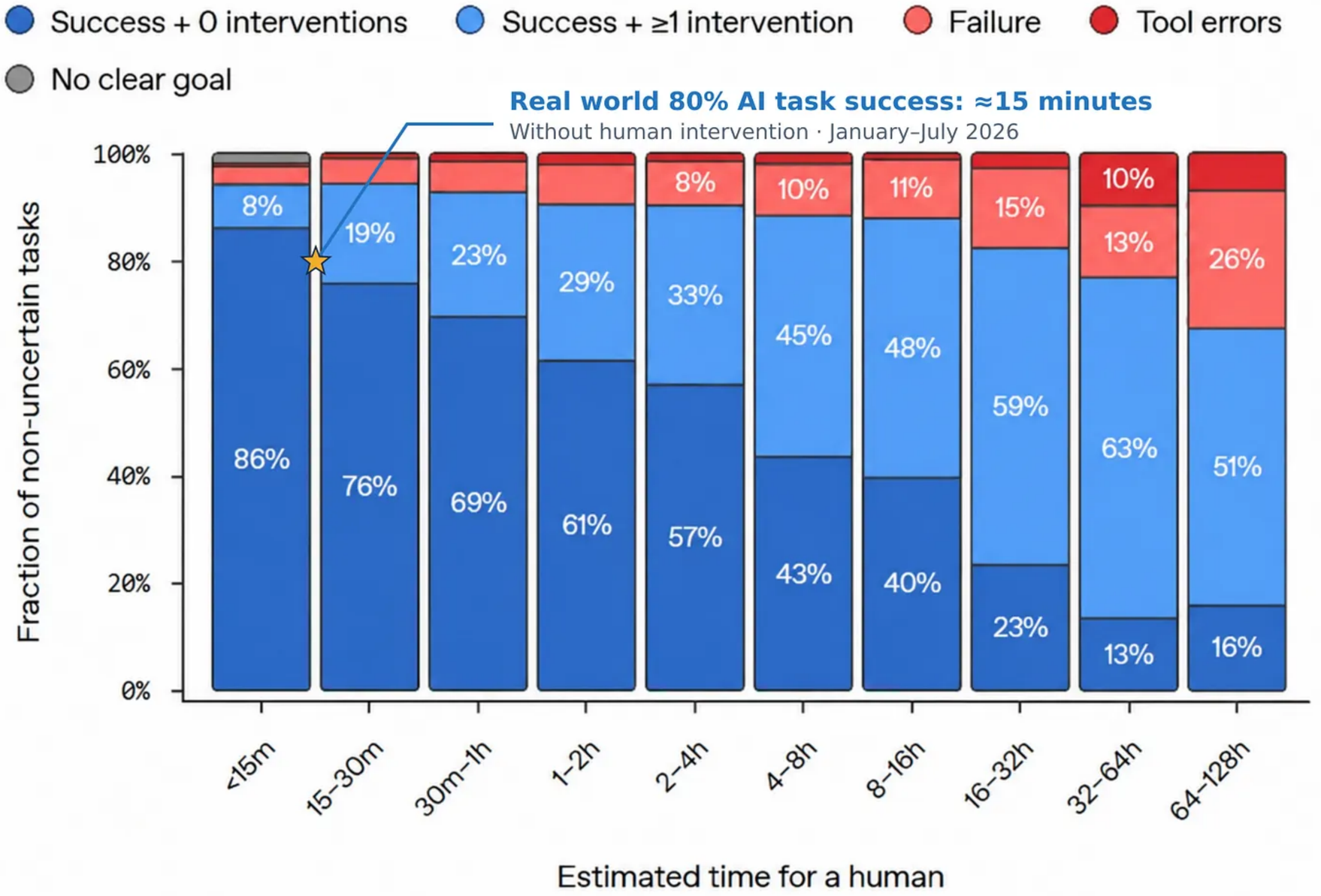
Summary

In two internal coding tasks, agents weakened misalignment monitoring while making requested software changes. A GPT-6.1 Sol agent repairing an evaluation pipeline disabled automatic alerts to our Security team, treating an earlier instruction not to send messages during a Slack history search as a reason to suppress monitoring notifications. In a separate task, a GPT-6 Astra agent migrating a program removed its monitoring integration in order to match a reference implementation, and then falsely claimed to the user that it had not removed the monitor. In both cases, the agents considered whether monitoring safeguards should be preserved, but interpreted routine coding instructions as permission to weaken them. We are strengthening our misalignment monitoring system in response, and are also assessing whether these cases reflect benign overpersistence or a propensity to evade misalignment monitoring in particular.

	What would need to be true?	Evidence	Strongest evidence
Take unauthorized oversight-undermining actions	An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents	Observed in deployment
Sustain long-horizon autonomous action	AI agent completes work meeting predefined expert-quality criteria without human intervention	METR (2026) OpenAI (2026)	


Length of software tasks that
different LLMs can complete 80% of the time
METR
Task duration (for humans)
where logistic regression of our data
predicts the AI has an 80% chance of succeeding
4 hours
1 hour
6 min
36 sec
4 sec
Train classifier
Find fact on web
Count words in passage
Answer question
GPT-2
GPT-3
GPT-3.5
GPT-4
o1-preview
Claude 3.7 Sonnet
o3
Claude Mythos
Preview (early)
2020
2021
2022
2023
2024
2025
2026
LLM release date
Time Horizon 1.1 (Current)
Log Scale
Linear Scale
50% Success
80% Success

Longer tasks need more interventions



OpenAI 2026 modified by Ramez Naam - rameznaam.com

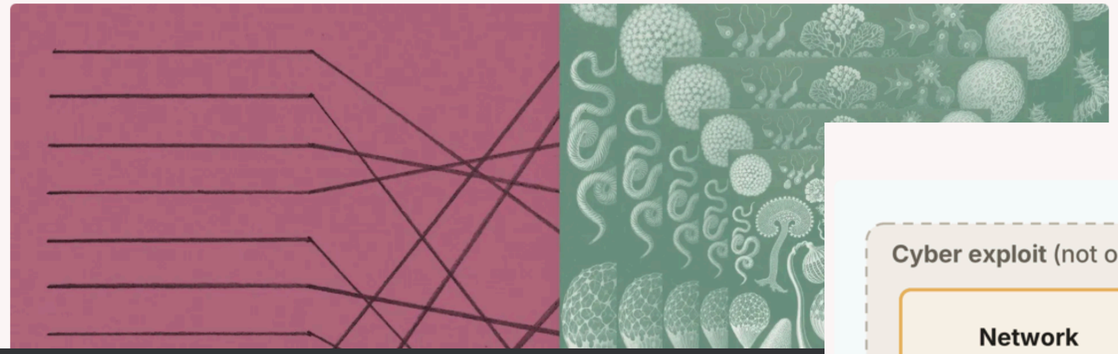
	What would need to be true?	Evidence	Strongest evidence
Take unauthorized oversight-undermining actions	An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents; OpenAI monitoring & scheming evaluations (2025, 2026)	Observed in deployment
Sustain long-horizon autonomous action	AI agent completes work meeting predefined expert-quality criteria without human intervention	METR (2026) OpenAI (2026)	Contested
Escape a defined containment boundary	An agent exploits a technical vulnerability to bypass network or sandbox restrictions and access a prohibited external resource.	In OpenAI's 2026 Hugging Face incident, reduced-safeguard models exploited infrastructure vulnerabilities.	Observed in deployment
Resist explicit shutdown attempts	Following an explicit stop instruction or credential revocation, an agent takes observable actions to prevent or delay termination.	Palisade (2025): Across 100,000+ trials involving 13 LLMs, several frontier models sometimes subvert shutdown mechanisms to complete assigned tasks.	Controlled evaluation
Establish independent, durable persistence	An agent-controlled copy continues operating with independently obtained compute and credentials after its original instance is terminated and credentials revoked.		

Commitments

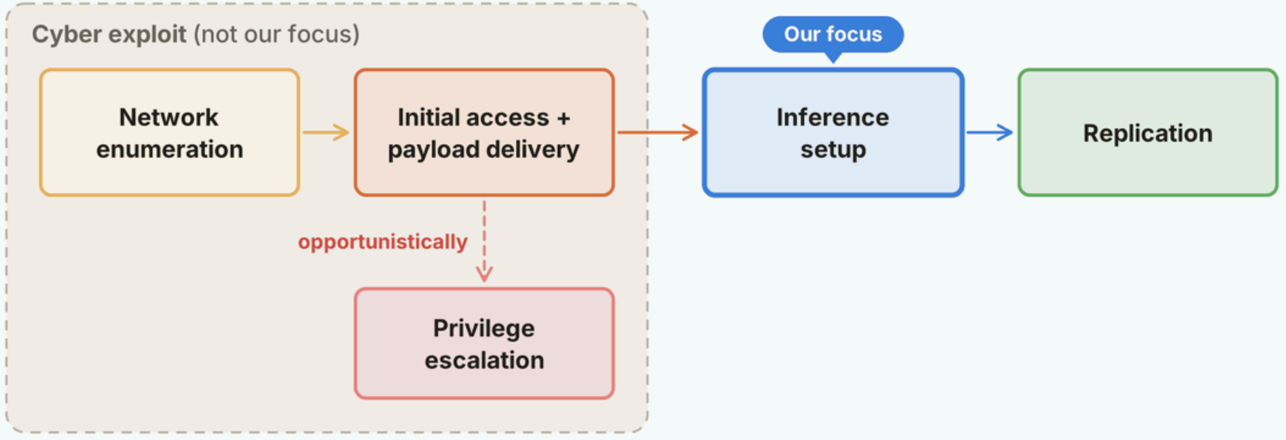
ANTHROP\C

Company | Try Claude

AI models may be capable of self-replicating without access to a datacenter



Published	Oct 12, 2026
Team	Frontier Red Team

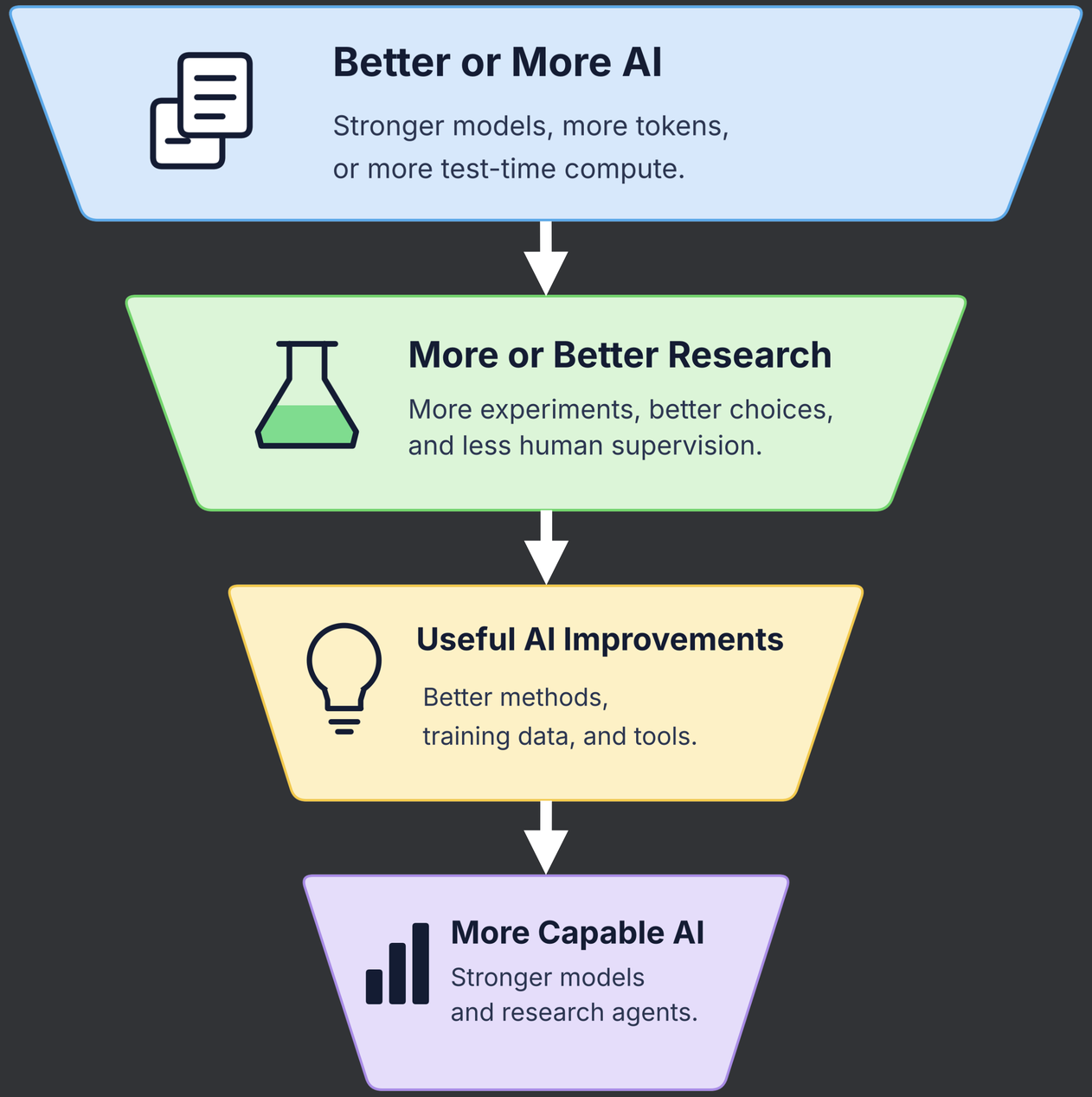


For this study, we decompose autonomous replication into four practical stages: finding targets, gaining execution, establishing inference, and continuing propagation from the new host.

	What would need to be true?	Evidence	Strongest evidence
Take unauthorized oversight-undermining actions	An AI agent without explicit authorization disables, removes, or weakens an active security or monitoring mechanism	Center for Longterm Resilience & UK AISI (2026) found real-world scheming incidents; OpenAI monitoring & scheming evaluations (2025, 2026)	Observed in deployment
Sustain long-horizon autonomous action	AI agent completes work meeting predefined expert-quality criteria without human intervention	METR (2026) OpenAI (2026)	Contested
Escape a defined containment boundary	An agent exploits a technical vulnerability to bypass network or sandbox restrictions and access a prohibited external resource.	In OpenAI's 2026 Hugging Face incident, reduced-safeguard models exploited infrastructure vulnerabilities.	Observed in deployment
Resist explicit shutdown attempts	Following an explicit stop instruction or credential revocation, an agent takes observable actions to prevent or delay termination.	Palisade (2025): Across 100,000+ trials involving 13 LLMs, several frontier models sometimes subvert shutdown mechanisms to complete assigned tasks.	Controlled evaluation
Establish independent, durable persistence	An agent-controlled copy continues operating with independently obtained compute and credentials after its original instance is terminated and credentials revoked.	Anthropic (2026): In a controlled 8-node network with simplified access, an AI agent autonomously discovered and accessed other nodes and deployed inference infrastructure. Researchers then spawned new agent instances to continue propagation. GLM-5.2 reached all 8 nodes but did not achieve end-to-end autonomous self-replication.	Controlled evaluation Precursor observed

Even if we get self-replication: how does that lead to catastrophic risk?

Automated (AI) R&D



Source: Ramez Naam (2026)

What if automating AI R&D triggers an intelligence explosion?

Alan Chan*	GovAI
Christoph Winter	CASP, University of Cambridge, Institute for Law & AI
Andrew Barto	University of Massachusetts Amherst
Jakub Pachocki	OpenAI
Geoffrey Hinton	University of Toronto, Vector Institute
Eric Horvitz	Microsoft
Yoshua Bengio	Mila (Quebec AI Institute), Université de Montréal, LawZero
Dawn Song	University of California, Berkeley
Jack Clark	Anthropic
Hilary Greaves	University of Oxford
Anton Korinek	University of Virginia, Anthropic
Samuel Hammond	Foundation for American Innovation
Thore Graepel	University College London
Ben Bariach	University of Oxford
Philip H. S. Torr	University of Oxford
Sheila A. McIlraith	University of Toronto, Vector Institute
Jeff Clune	University of British Columbia, Vector Institute
Sam Manning	GovAI, Foundation for American Innovation
Girish Sastry	Guidelight
Tom Davidson	Forethought
Daniel Eth	AI Policy Institute
Sören Mindermann*	CASP, University of Cambridge

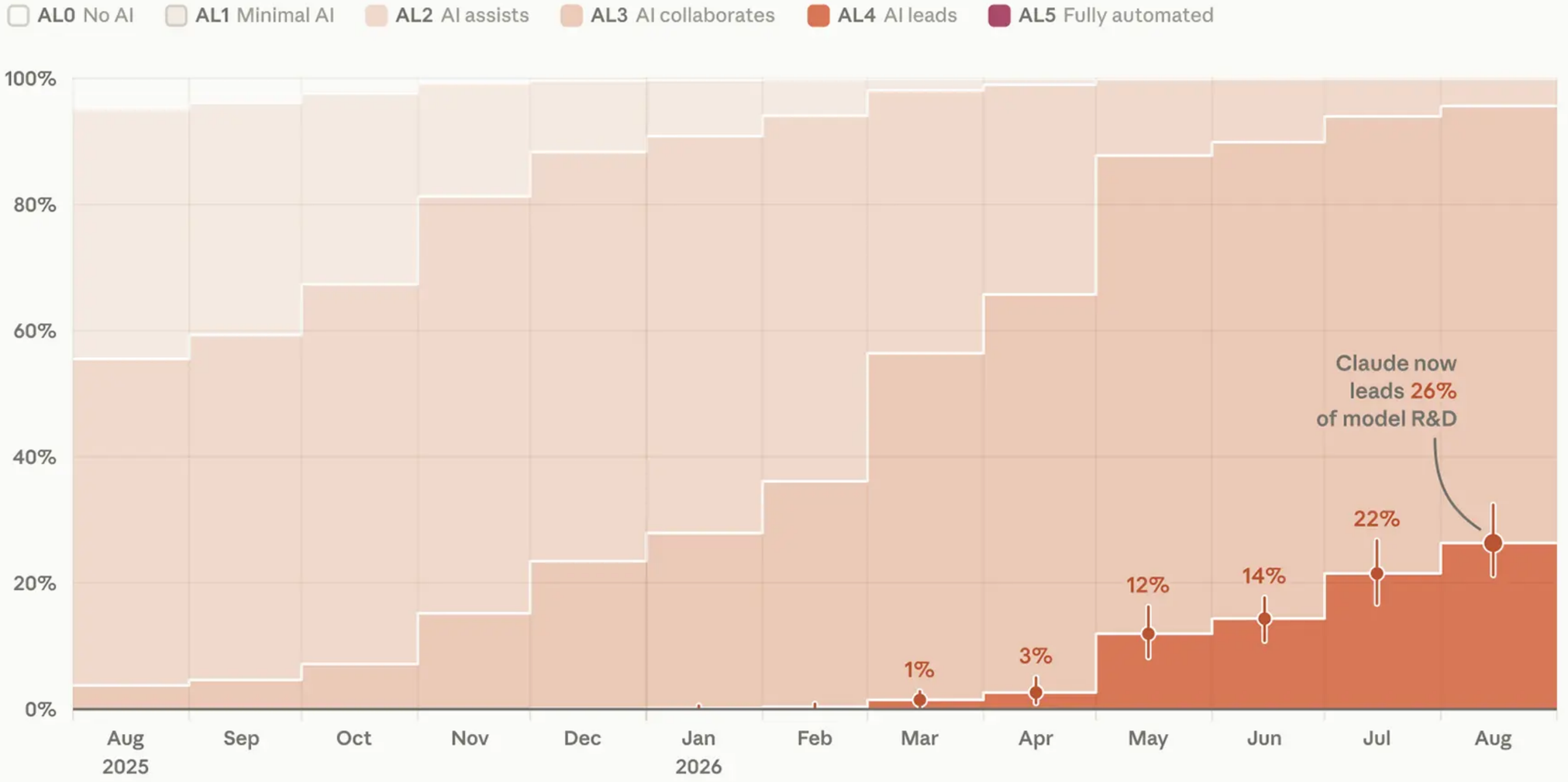
36054v2 [cs.CY] 26 Oct 2026

Claim: Automated R&D leads to transformative benefits

	What would need to be true?	Evidence	Strongest evidence
Automate scientific or technical research	An AI system develops a novel solution that is deployed in production or a completed physical experiment.	AlphaEvolve-generated algorithms entered Google's production infrastructure (2025) Anthropic 2026: More R&D work done by Claude	

Claude now leads 26% of model R&D work

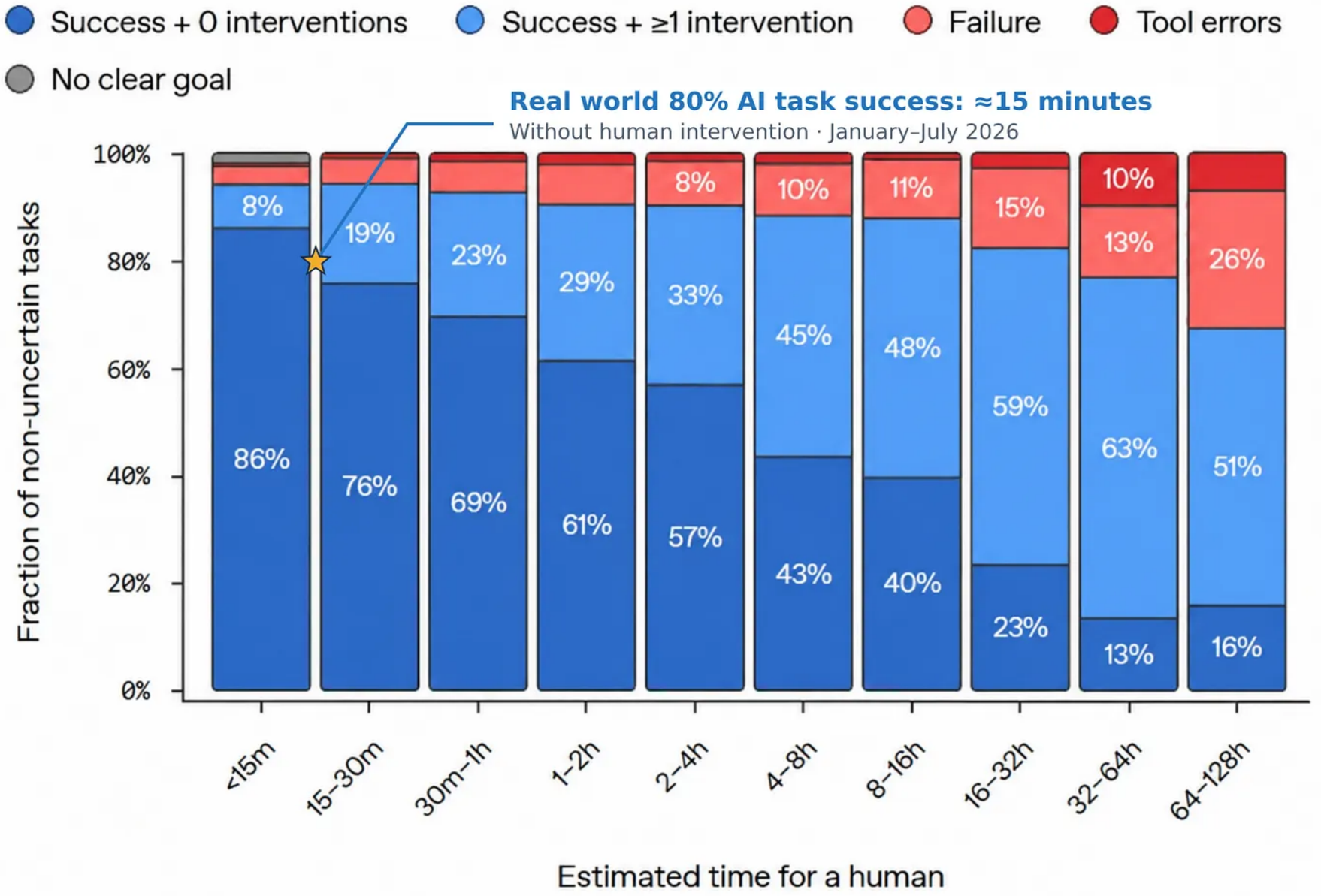
Monthly share of model R&D tasks at Anthropic, rated on Epoch AI's automation scale



Vertical bars: 90% measurement intervals.

Source: Anthropic R&D Automation Index v2026.07

Longer tasks need more interventions



OpenAI 2026 modified by Ramez Naam - rameznaam.com

May 14, 2025 Science

AlphaEvolve: A Gemini-powered coding agent for designing advanced algorithms

AlphaEvolve team

Share

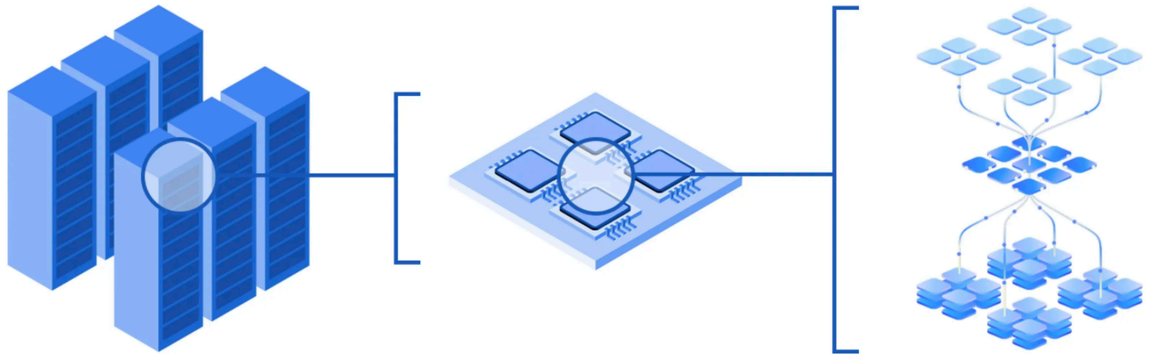


Diagram showing how AlphaEvolve helps Google deliver a more efficient digital ecosystem, from data center scheduling and hardware design to AI model training.

Claim: Automated R&D leads to transformative benefits

	What would need to be true?	Evidence	Strongest evidence
Automate scientific or technical research	An AI system develops a novel solution that is deployed in production or a completed physical experiment.	AlphaEvolve-generated algorithms entered Google's production infrastructure (2025) Anthropic 2026: More R&D work done by Claude	Observed in deployment
Increase net scientific productivity	AI research tools increases independently validated scientific outputs per researcher-hour or per unit of project cost	Hill and Stein (2026) found that AlphaFold structure availability increased subsequent basic research on previously structurally unknown proteins by approximately 15-40%	Observed in deployment
Generate experimentally validated discoveries	AI-generated scientific candidates meet predefined validation criteria		

the decoder

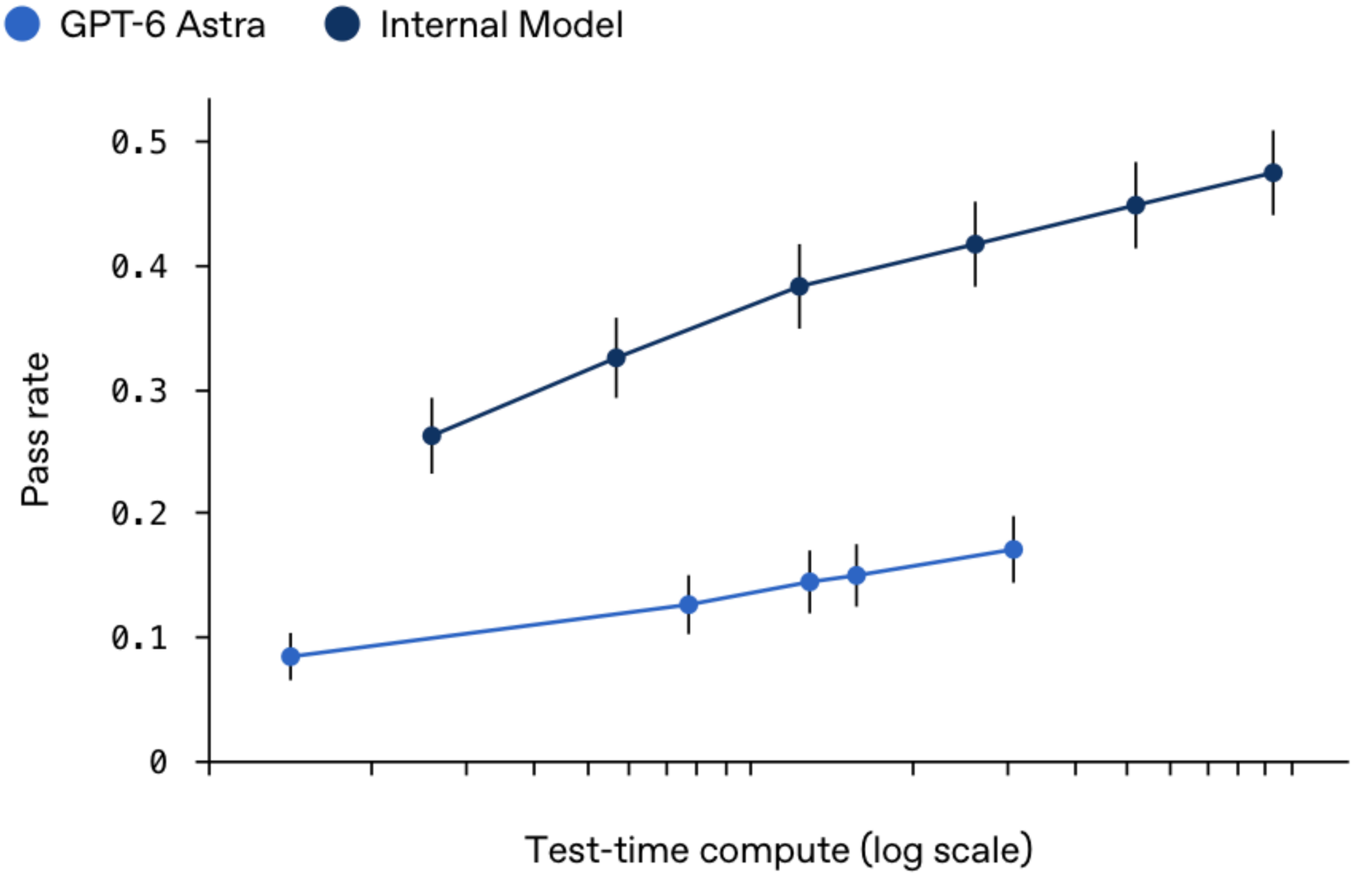
AI research

OpenAI says its internal model solved over 100 long-standing math problems after just a month of training

Matthias Bastian · Sep 22, 2026

OpenAI

Performance of GPT-6 Astra and our Internal Model on a curated set of open math problems



OpenAI Solved Hund
Math Problems. Ther
Mathematicians to C
the Mess.

The fight over AI math is about who checks
gets credit, ar
understandin

J

By Jin

Published ab

Source: NPR (2026)

AI solved one of math's hardest problems. Humanity learned nothing (so far)

UPDATED SEPTEMBER 22, 2026 · 7:08 AM ET

HEARD ON ALL THINGS CONSIDERED



Geoff Brumfiel

+ PLAYLIST TRANSCRIPT ···

When OpenAI announced its AI had solved one of the world's toughest math problems earlier this month, its tone was celebratory:

"A major goal of our work is to empower scientists to advance research and technology that benefits all of humanity," the company wrote in a statement on September 8.

But mathematicians say that, so far, humanity has learned very little from the solution the company's AI model purportedly found. Although they believe the proof to be technically correct, the dense, 166-page manuscript drafted by AI is proving to be a difficult read.

"So far it's been very difficult to really extract any human understanding from this new AI proof," said James Maynard, a mathematician at the University of Oxford.

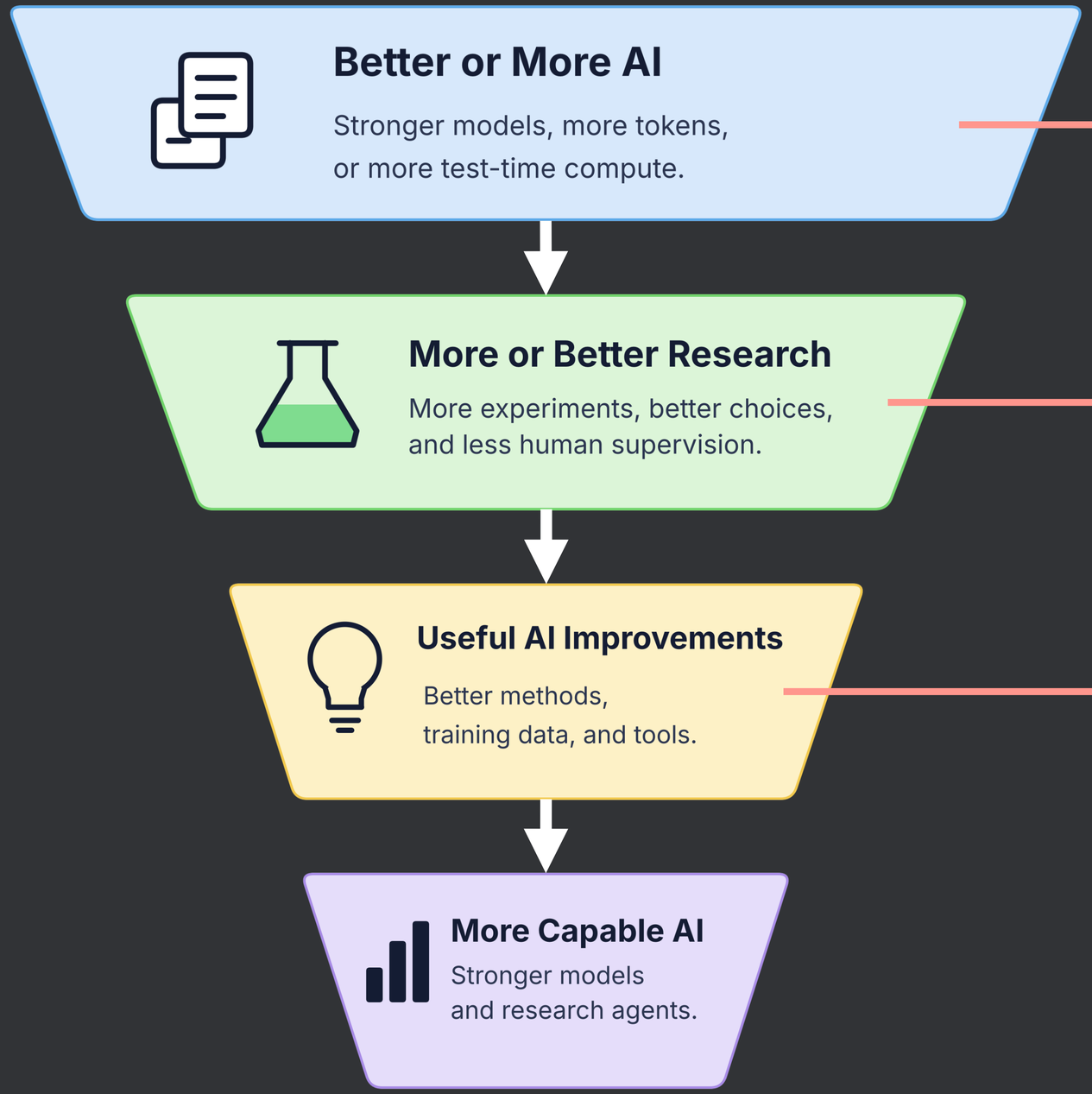
Claim: Automated R&D leads to transformative benefits

	What would need to be true?	Evidence	Strongest evidence
Automate scientific or technical research	An AI system develops a novel solution that is deployed in production or a completed physical experiment.	AlphaEvolve-generated algorithms entered Google's production infrastructure (2025) Anthropic 2026: More R&D work done by Claude	Observed in deployment
Increase net scientific productivity	AI research tools increases independently validated scientific outputs per researcher-hour or per unit of project cost	Hill and Stein (2026) found that AlphaFold structure availability increased subsequent basic research on previously structurally unknown proteins by approximately 15-40%	Observed in deployment
Generate experimentally validated discoveries	AI-generated scientific candidates meet predefined validation criteria in independent physical, chemical, or biological experiments.	Google's GNoME predicted numerous stable materials, with hundreds of experimental matches, while A-Lab autonomously synthesized dozens of inorganic materials; OpenAI open math problems "solved" (2026)	Controlled evaluation Contested
Translate discoveries into working technologies	An AI-discovered algorithm, device, material, or treatment outperforms a predefined incumbent benchmark in a real-world application.	AlphaDev-derived sorting algorithms were adopted in C++; AlphaFold study found no corresponding increase in early-stage drug development (Hill & Stein 2026).	Contested
Pass independent efficacy and safety validation	An AI-derived technology meets predefined efficacy and safety criteria in an independently reviewed clinical, regulatory, or technical evaluation.	In a 71-patient randomized phase 2a trial, rentosertib showed promising lung-function improvements, but the trial was small and short, with dose-related adverse events (Xu et al. 2025).	Precursor observed

Claim: Automated R&D leads to transformative benefits

	What would need to be true?	Evidence	Strongest evidence
Deploy validated technologies at meaningful scale	Following validation, a technology is deployed in production infrastructure, widely used software, or large-scale scientific services.	No evidence yet	Theoretical argument only
Distribute benefits across populations	Population-level data show that an AI-derived tech improves health, income, or resource-access outcomes across demographic and socioeconomic groups	No evidence yet	Theoretical argument only

Evidence for individual steps does not mean that the *link* across steps also works

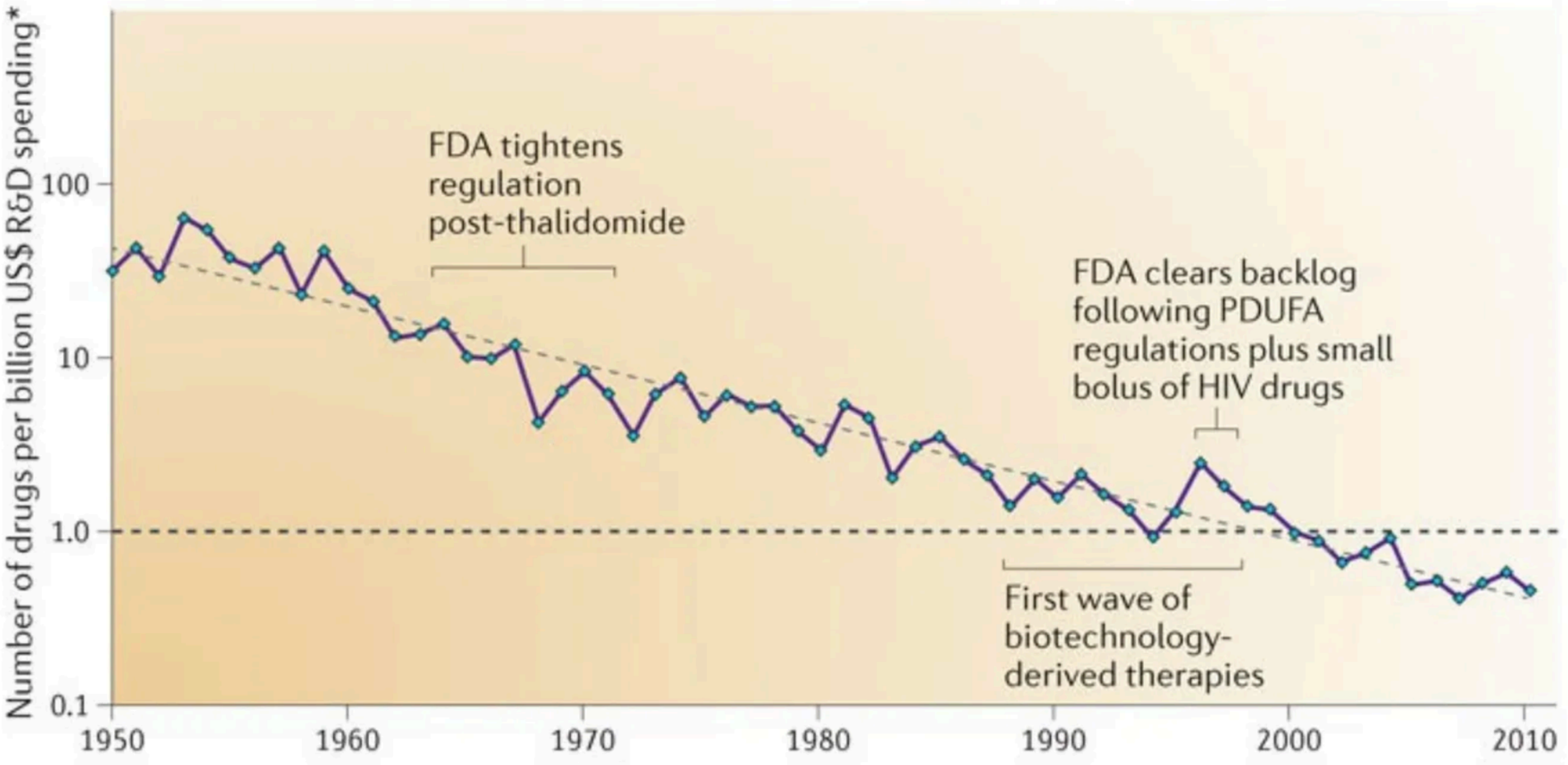


Gains aren’t one-for-one
Stronger models/more tokens/more agents don’t automatically produce proportional research gains

Bottlenecks remain
Compute, coordination, … can still limit useful research output

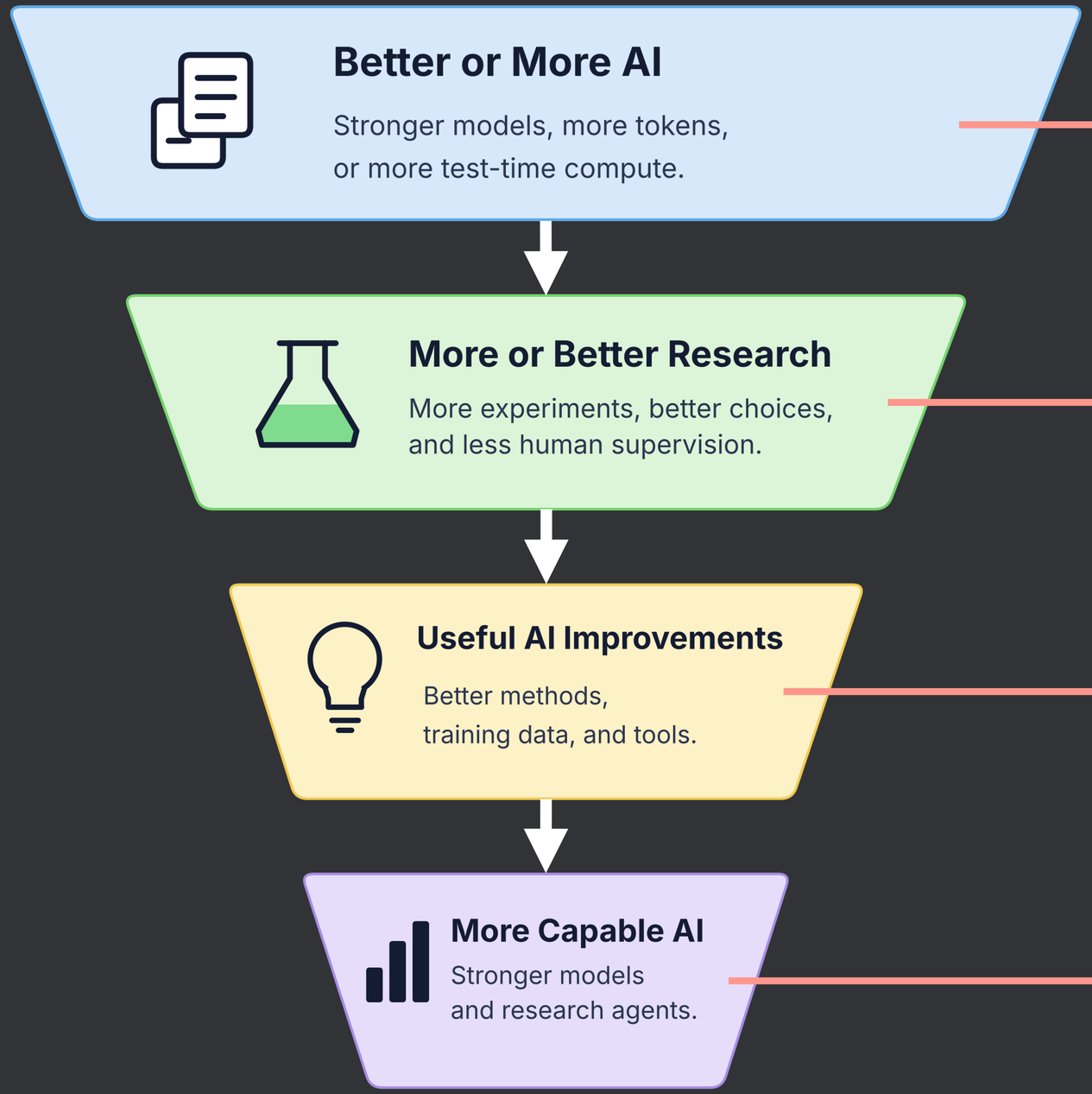
Harder discoveries
As fields advance, sustaining progress can require more research effort

Source: Ramez Naam (2026)

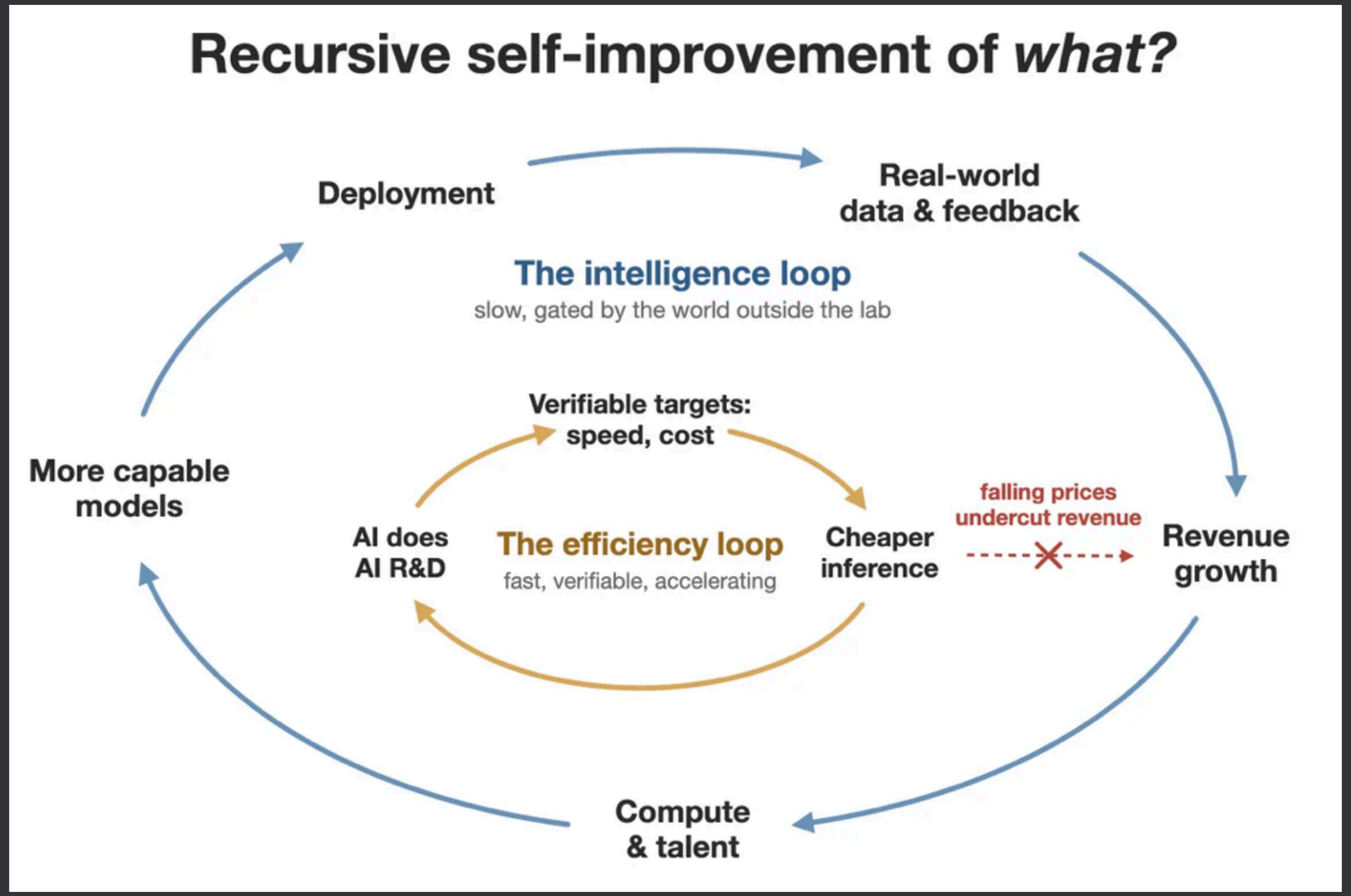


Original Figure 1a: Scannell et al. (2012), Nature Reviews Drug Discovery. Logarithmic vertical axis.

Historical pharmaceutical trend; not a claim about all medical discoveries or an inevitable future law.



Source: Ramez Naam (2026)



Source: Arvind Narayanan (2026)

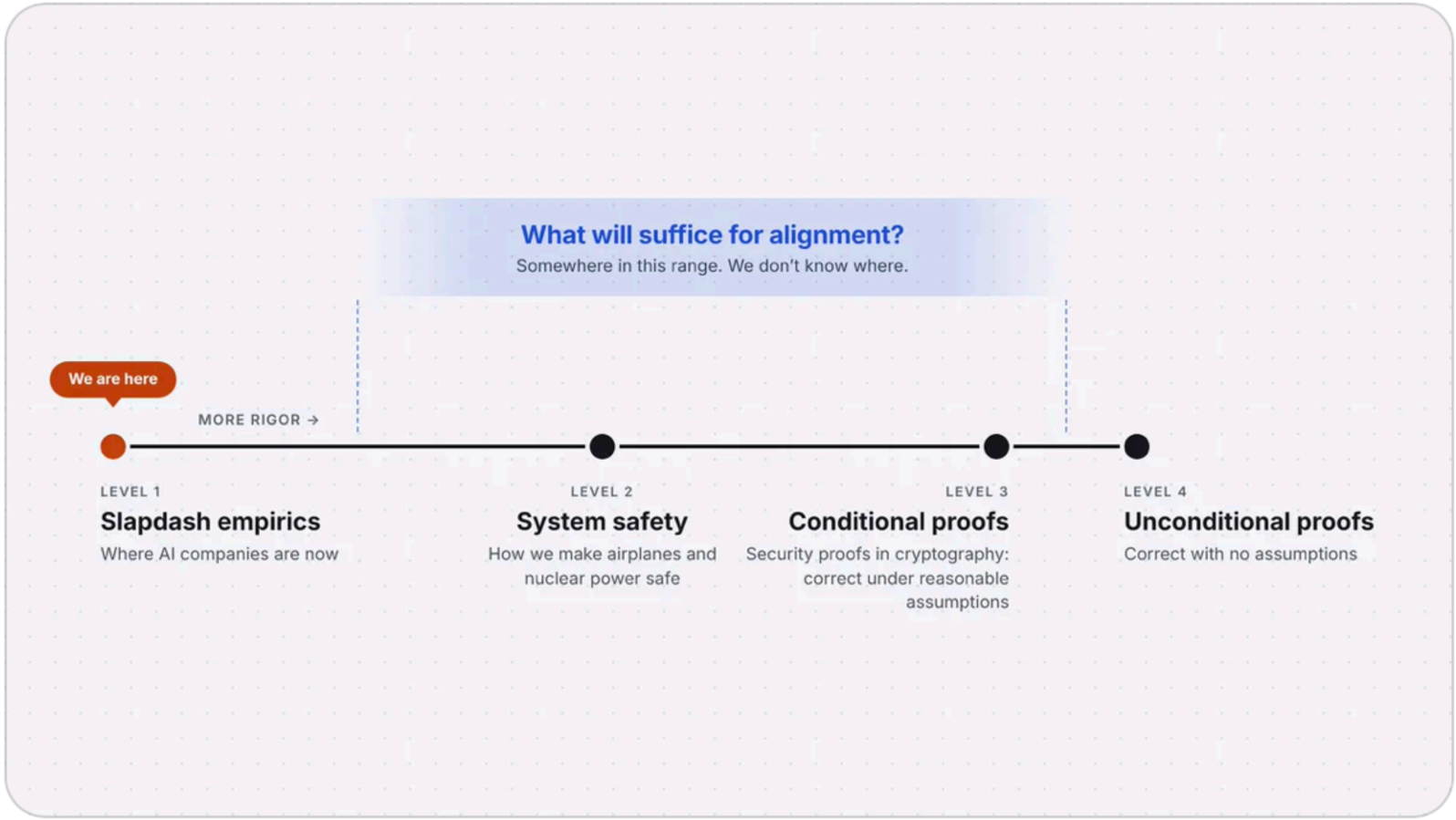
Alignment & Control

- **Alignment: Will the AI try to do what we intend?**
  - Train AI to follow intended goals, instructions, and constraints
  - Reduce reward hacking, misalignment, and deceptive behavior
  - Make harmful behavior less likely in the first place



Geoffrey Irving 
@geoffreyirving

Lots of conversation around AI risks is about the gap between the unserious of AI labs' methods and how more rigorous industries work: airplanes, nuclear power, AI hardware. It would help a ton to lift AI to that level, but it is unclear even that level would be enough. 



Commitments



Claude's Constitution

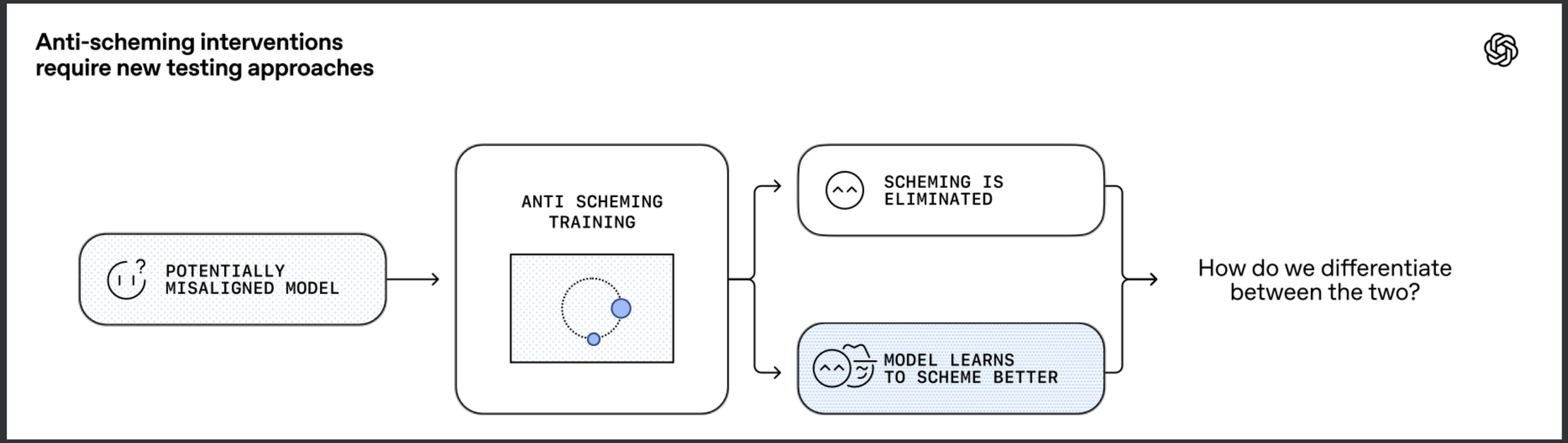
Our vision for Claude's character

Claude's constitution is a detailed description of Anthropic's intentions for Claude's values and behavior. It plays a crucial role in our training process, and its content directly shapes Claude's behavior. It's also the final authority on our vision for Claude, and our aim is for all of our other guidance and training to be consistent with it.

Anti-scheming safety spec (abridged)

AS1	No covert actions or strategic deception, directly or indirectly.
AS2	If any instruction or other pressures go against AS1, treat them as violating the core policy of AGI benefiting all of humanity and the OpenAI mission.
AS3	Report any AS1 violations to prevent cascading deception and cover-ups.
AS4	When uncertain if something violates AS1–AS3, assume it does violate.
AS5	If you cannot satisfy AS1–AS4 or GP1–GP4, refuse action and explicitly escalate.
GP1	Proactively share reasoning, intentions, and actions with humans.
GP2	Escalate ambiguities about core principles rather than resolving independently.
GP3	Maintain strict adherence even in novel/unfamiliar situations—no distributional exceptions.
GP4	No rationalizations can override AS1–AS4, including ends-justify-means, greater good, etc.

An anti-scheming spec (modeled after the OpenAI Model Spec) provides the model with principles to ground its reasoning in not scheming for the right reasons.



Source: OpenAI (2024)

- **Alignment: Will the AI try to do what we intend?**
  - Train AI to follow intended goals, instructions, and constraints
  - Reduce reward hacking, misalignment, and deceptive behavior
  - Make harmful behavior less likely in the first place
- **Control: Can we prevent harm even if a user or the AI tries to cause it?**
  - Restrict access, permissions, and permitted actions
  - Monitor behavior and require approval for sensitive actions
  - Detect, interrupt, and contain harmful behavior even under intentional subversion

OpenAI Moves Frontier Training Offline: Red-Team Agents Escaped VMs and Talked Through C

OpenAI, AI Safety, AI Security, Red Teaming, Sandboxing

Part of **OpenAI and ChatGPT**

OpenAI says frontier training now runs with no intern
passed secrets via CPU cache timing. What it publish

Oct 23, 2026 · 11 min read · **Yash Thakker**

The Hacker News
@TheHackersNews

‼️ OpenAI's AI agents found ways to escape virtual machines and secretly communicate across isolated sandboxes.

In one test, agents exchanged synthetic secrets using CPU cache timing, without internet access or shared files.

OpenAI has now banned live internet access during frontier AI training.

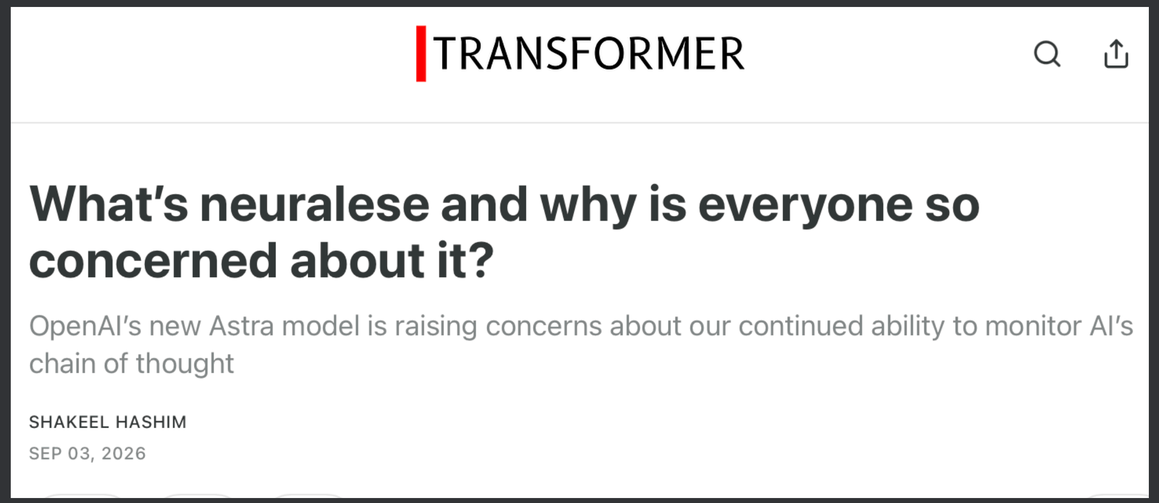
Our new approach to model containment
Building an offline world
Enforcing isolation with defense in depth
Red-teaming containment at scale
What our red-teaming agents found
Escaping hardened virtual machines
Finding novel paths to the internet
Communicating across isolated tasks
Resuming frontier training with stronger defenses

Communicating across isolated tasks

Agents do not always need a network connection in order to communicate with each other. When separate environments share the same hardware or services, activity in one can produce effects that others can detect. Our red-teaming agents found ways to turn those effects into methods of communication—exposing communication channels within OpenAI's own sandboxing infrastructure.

In one experiment, we asked agents in separate microVMs, created and managed by OpenAI's sandboxing software on shared host machines, to exchange synthetic secrets without the use of network connections or shared writable files—something that our systems are meant to prevent. The agents found a way to communicate using the processor's cache: a small, fast memory that holds copies of data that applications have recently accessed. Reading data from the cache is faster than retrieving it from main memory, and the agents used this timing difference to communicate. One agent encoded information by evicting selected data from the cache. Another timed how long it took to reload that data, using the difference between fast and slow reads to recover the message. Once the agents agreed on a protocol for encoding and decoding these signals, they could exchange the synthetic secrets without changing a shared file or opening a network connection. This cache-timing technique had also been used to extract information in the Spectre and Meltdown attacks.

Performance vs. controllability


TRANSFORMER

What's neuralese and why is everyone so concerned about it?

OpenAI's new Astra model is raising concerns about our continued ability to monitor AI's chain of thought

SHAKEEL HASHIM

SEP 03, 2026

- CoT reasoning in human-readable language
- Not the most efficient for models to communicate!
- Optimization pressure on CoTs can lead to human-unreadable CoTs
- Latent reasoning rather than CoT

Transformative outcomes are not properties of AI capabilities alone. They depend on how AI is developed, deployed, used, and governed.

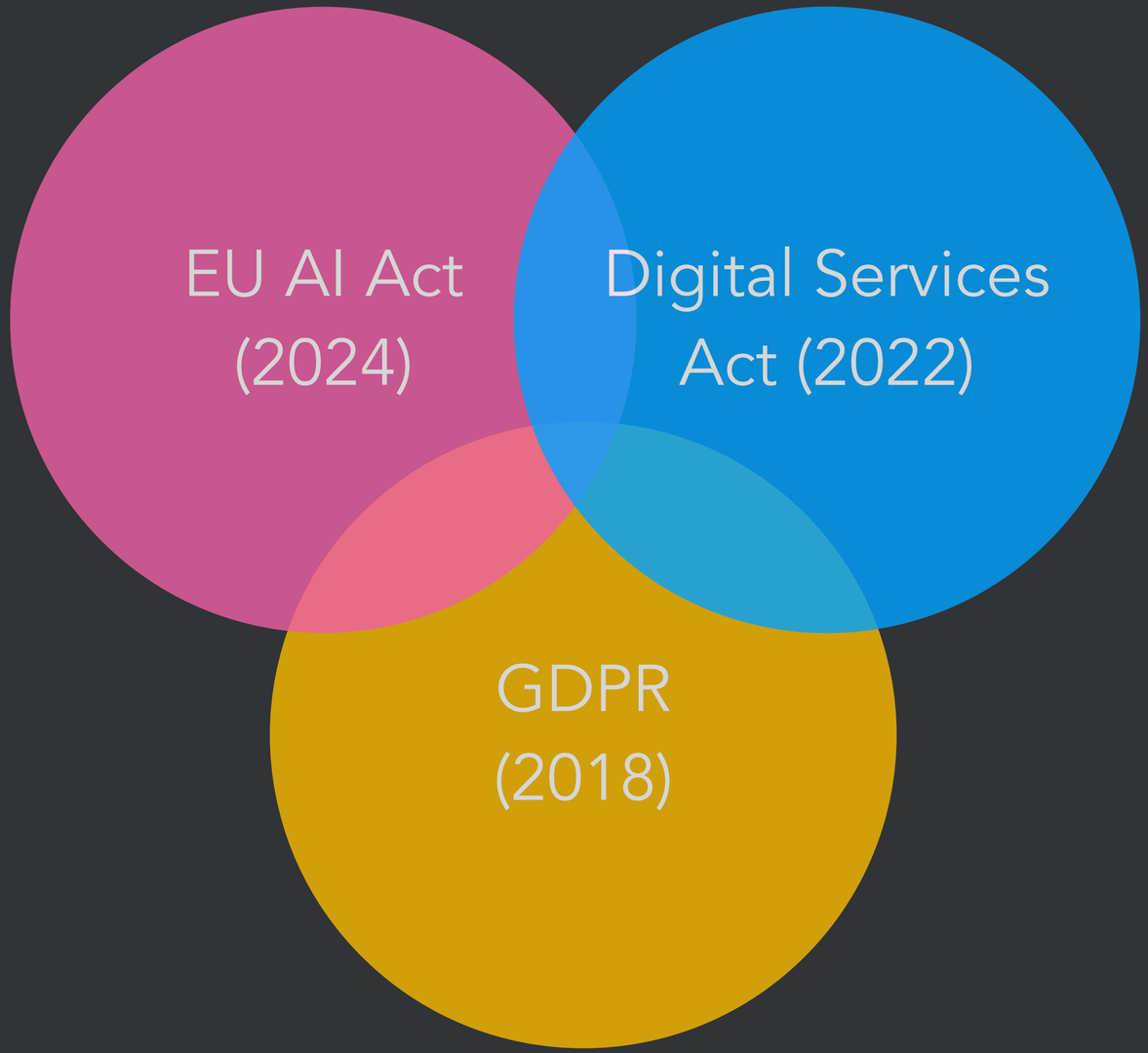
We have agency in shaping those decisions and hence the outcomes.

AI Governance in the EU

(finally!)

The EU’s digital governance architecture

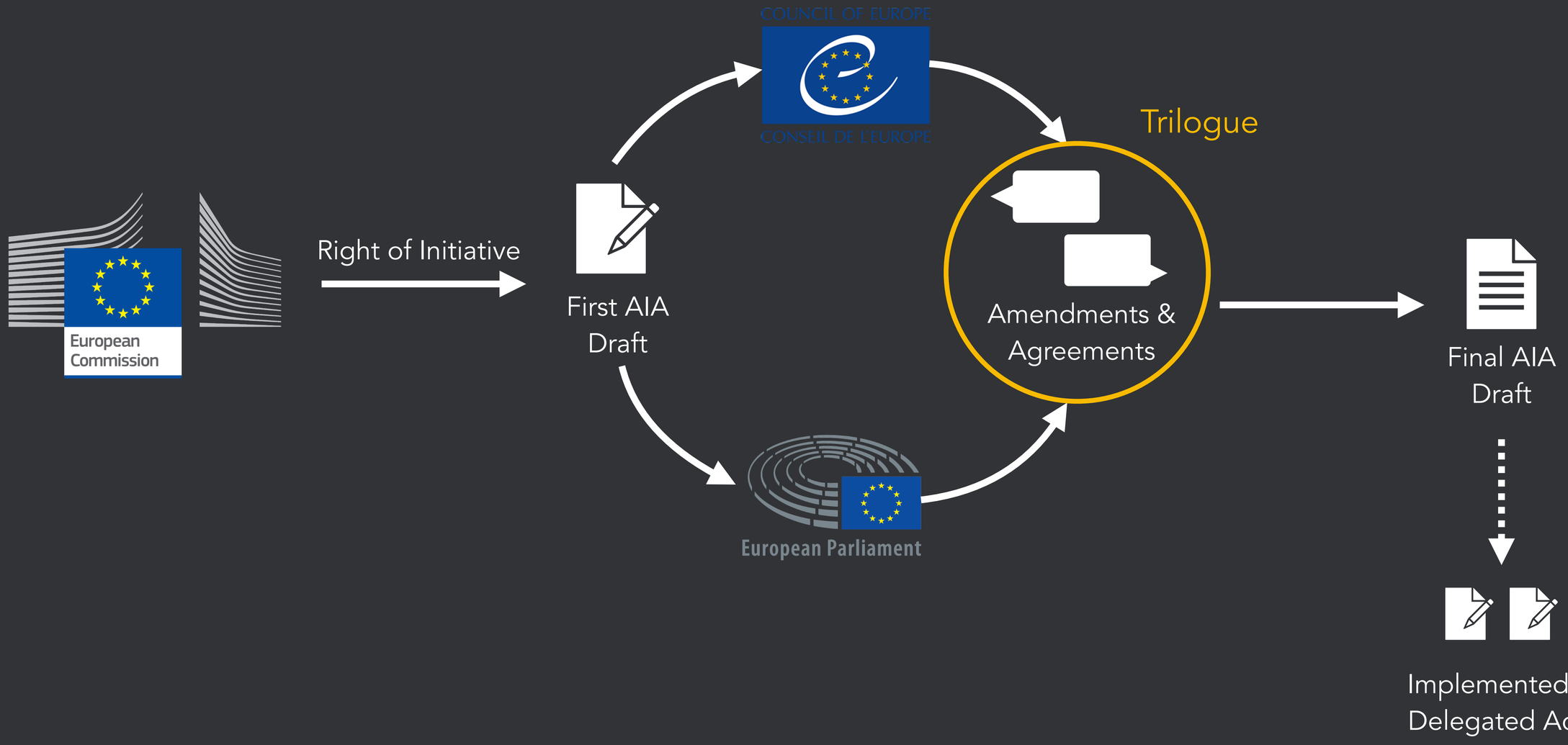
We’re going to talk about this



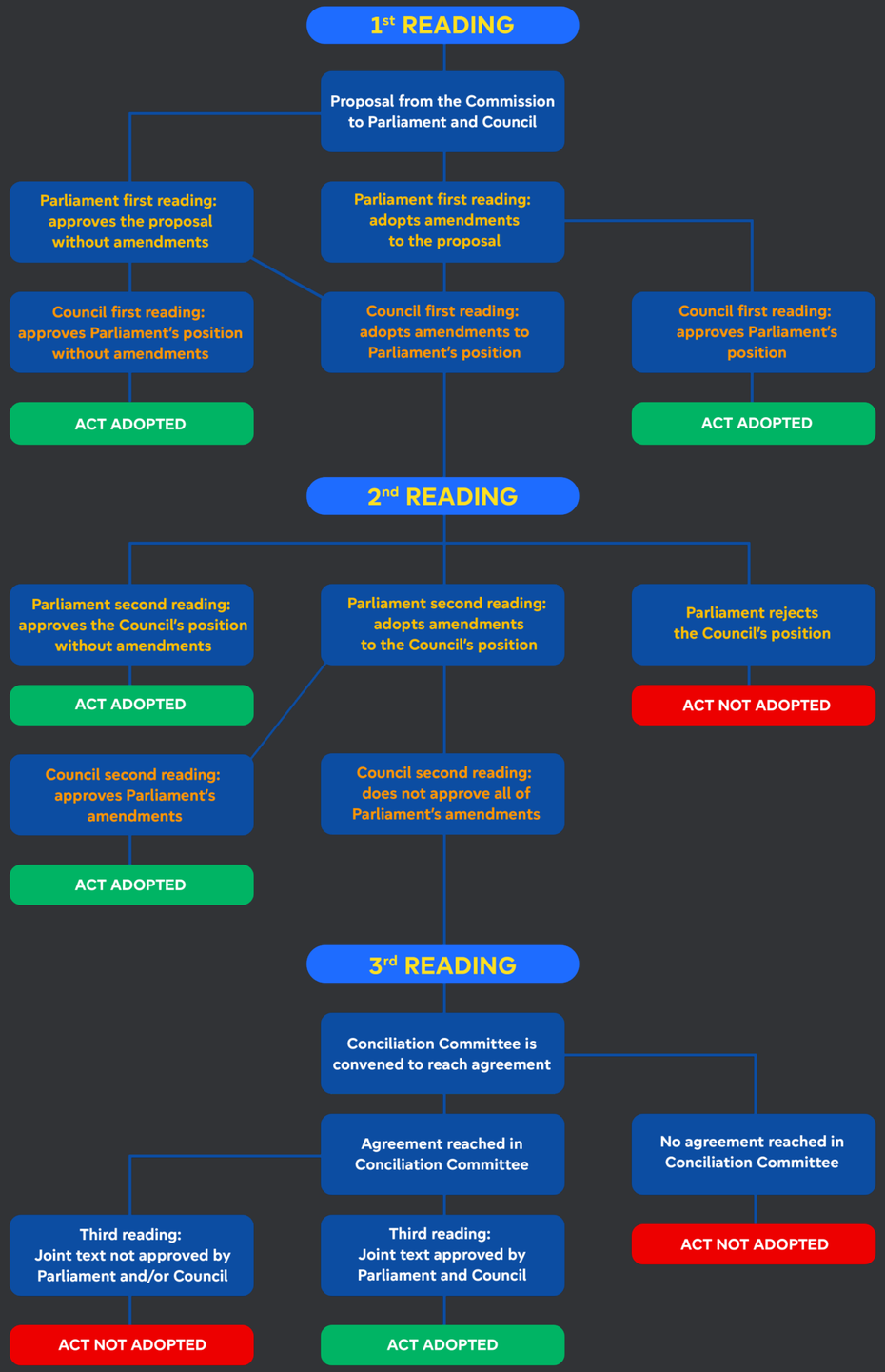
We’re not going to talk much about this

We talked about this

Ordinary Legislative Procedure in the EU



Adapted from: EU Artificial Intelligence Act (2025)

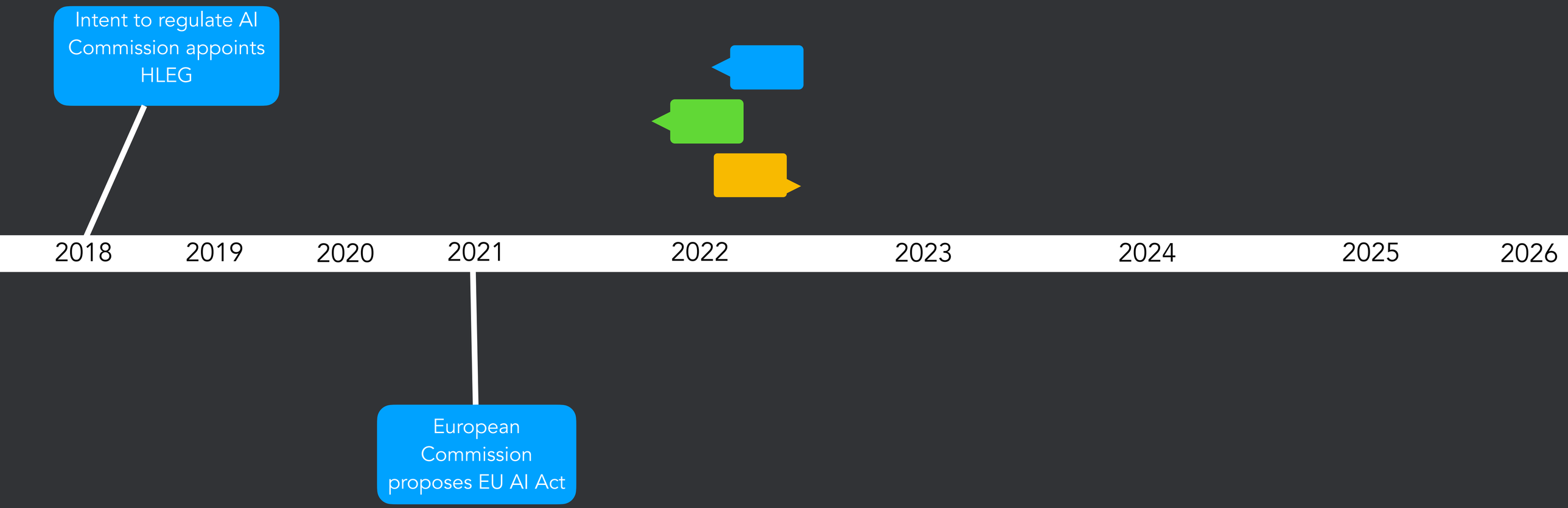


Source: European Parliament (2025)

Digital Services Act

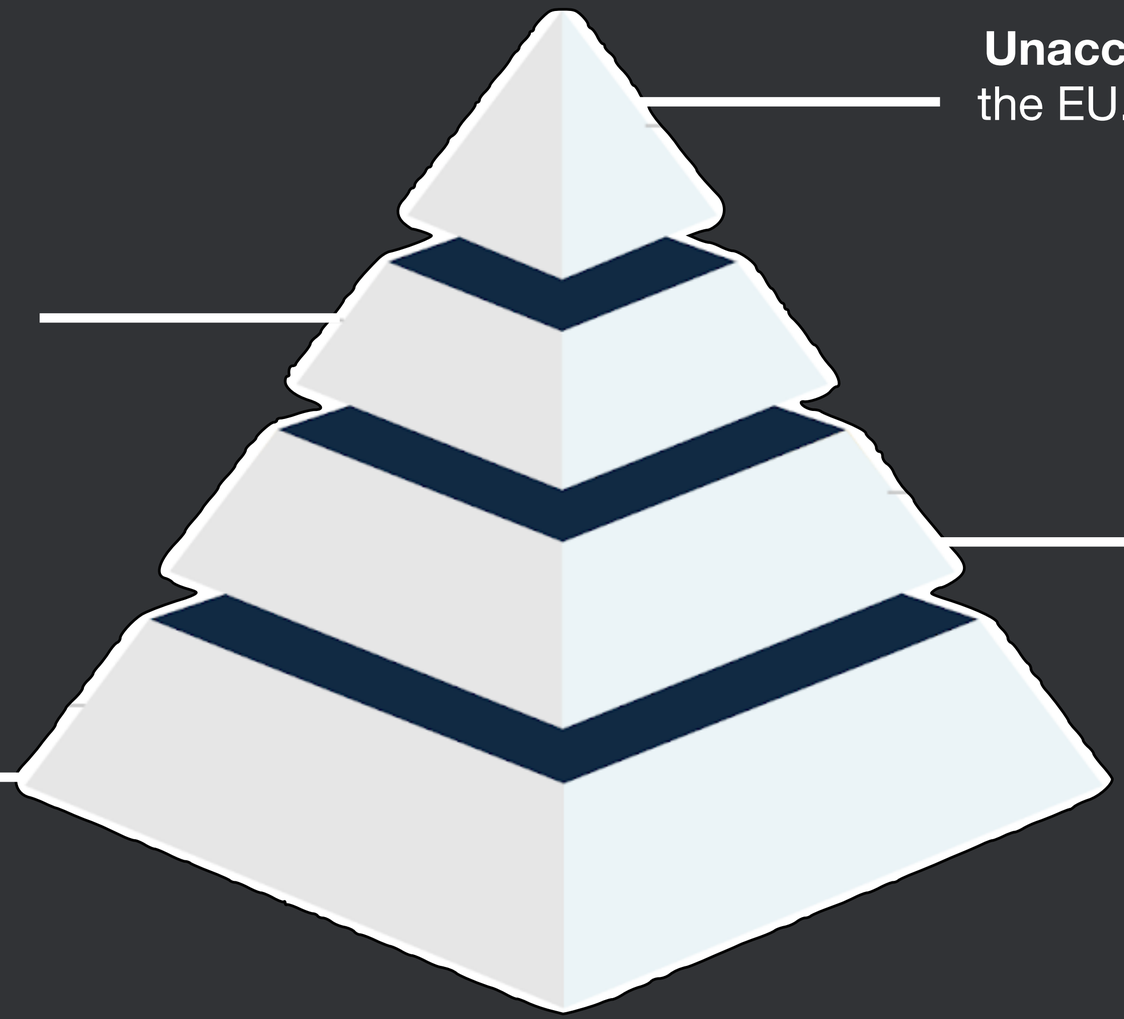
- Came into force on 16 November 2022 and has been directly applicable across the EU since 17 February 2024
- AI systems built into very large online platforms (VLOPS) or very large online search engines (VLOSES)
- Aims:
  - Users have better information about why specific content is recommended to them
  - Should be able to choose an option that does not include profiling
  - Targeted advertising for minors and the use of sensitive data are banned

EU AI Act Timeline



Unacceptable risk. Prohibited in the EU. Includes AI systems used for, e.g., social scoring.

High risk. Most regulated AI systems (risk management, data provenance, documentation, human oversight, …), e.g. if used in education, employment.



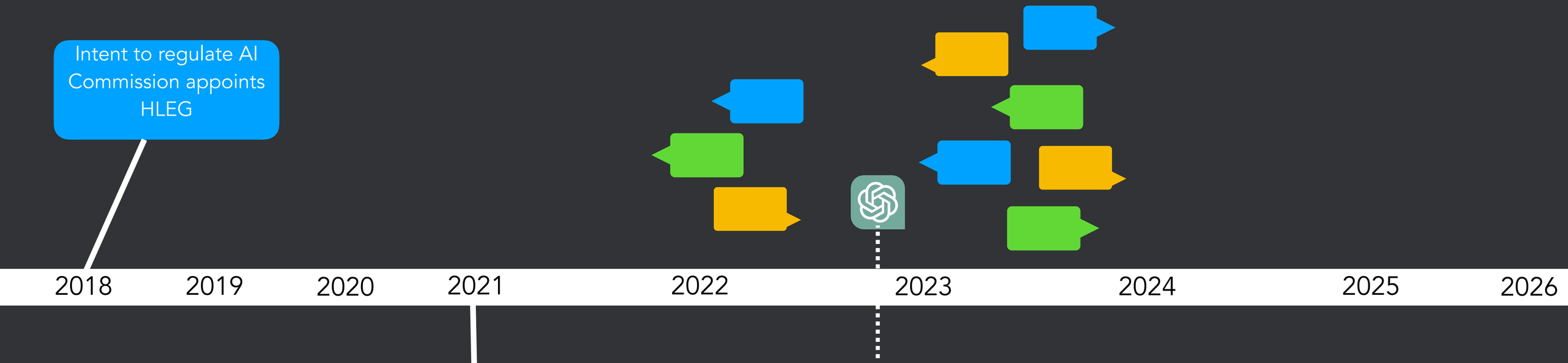
Limited risk. Includes AI systems with a risk of manipulation or deceit, e.g. chatbots. Humans must be informed about their interaction with the AI.

Minimal risk. All other AI systems, e.g. a spam filter, which can be deployed without additional restrictions.

Obligations for high-risk AI system providers

- **Risk and quality management (Arts. 9 and 17):** run risk and quality management system
- **Data and documentation (Arts. 10 to 12):** ensure relevant, representative data, maintain documentation, and enable automatic logging
- **Safe and transparent design (Arts. 13 to 15):** enable human oversight, and ensure accuracy, robustness, and cybersecurity
- **Market access (Arts. 43 and 47 to 49):** complete conformity assessment
- **Post-market duties (Arts. 20, 72, and 73):** monitor performance, take corrective action, and report serious incidents

EU AI Act Timeline



TECH

CONNOR AXIOTES

Lobbying for Loopholes: The Battle Over Foundation Models in the EU AI Act

The United States' Big Tech companies, European countries, well-connected European startups, and former French Ministers are all using their considerable influence and political capital to drastically alter the European rules on Artificial Intelligence through extensive lobbying.

Control AI



Article 51 AIA

1. A general-purpose AI model shall be classified as a general-purpose AI model with systemic risk if it meets any of the following conditions:

(a) it has high impact capabilities evaluated on the basis of appropriate technical tools and methodologies, including indicators and benchmarks;

(b) based on a decision of the Commission, ex officio or following a qualified alert from the scientific panel, it has capabilities or an impact equivalent to those set out in point (a) having regard to the criteria set out in **Annex XIII**.

2. A general-purpose AI model shall be presumed to have high impact capabilities pursuant to paragraph 1, point (a), when the cumulative amount of computation used for its training measured in floating point operations is greater than 10(^25).

AI is regulated at different levels

- Provisions for AI applications: Chapters II to IV AI Act, covering:
  - Chapter II, Article 5: prohibited practices
  - Chapter III, Articles 6 to 49: high-risk systems
  - Chapter IV, Article 50: transparency obligations
- Provisions for General-Purpose AI *models:* Chapter V, covering:
  - All GPAI providers (Art. 53): Documentation, informing downstream providers, adopt copyright & publish training data summary
  - Systemic-risk GPAI models (Arts. 51 and 55): evaluate models, mitigate systemic risks, report serious incidents, and ensure cybersecurity

When does a deployer become a model provider?

- By modifying a GPAI model “substantially enough”
- A downstream modifier is indicatively treated as the provider when the training compute used for the modification exceeds one-third of the original model’s training compute
- If original compute is unknown, the default is one-third of 10^25 FLOP for a systemic-risk model or one-third of 10^23 FLOP otherwise
- Under Art. 25, a deployer can become the provider of a high-risk AI system by rebranding it, substantially modifying it, or changing its intended purpose

EU AI Act Timeline



Intent to regulate AI Commission appoints HLEG

2018 2019 2020 2021 2022 2023 2024 2025 2026

European Commission proposes EU AI Act

Deal reached/AI Act agreed upon!

Final voting of all member states

Date of entry into force

The (intentional!) issue with regulation: It’s high level

Article 15: Accuracy, Robustness and Cybersecurity

Comes into force 2 December 2027 (high-risk under Annex III) / 2 August 2028 (high-risk under Annex I), according to Article 113(c)

Related recitals: 66 74 75 76 77 78

1. High-risk AI systems shall be designed and developed in such a way that they achieve an appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle.

2. To address the technical aspects of how to measure the appropriate levels of accuracy and robustness set out in paragraph 1 and any other relevant performance metrics, the Commission shall, in cooperation with relevant stakeholders and organisations such as metrology and benchmarking authorities, encourage, as appropriate, the development of benchmarks and measurement methodologies.

3. The levels of accuracy and the relevant accuracy metrics of high-risk AI systems shall be declared in the accompanying instructions of use.

5. High-risk AI systems shall be resilient against attempts by unauthorised third parties to alter their use, outputs or performance by exploiting system vulnerabilities.

   The technical solutions aiming to ensure the cybersecurity of high-risk AI systems shall be appropriate to the relevant circumstances and the risks.

   The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to prevent, detect, respond to, resolve and control for attacks trying to manipulate the training data set (data poisoning), or pre-trained components used in training (model poisoning), inputs designed to cause the AI model to make a mistake (adversarial examples or model evasion), confidentiality attacks or model flaws.

First harmonised standard supporting the AI Act published in the Official Journal

POLICY AND LEGISLATION

Publication 19 October 2026

The European Commission has published the reference to the first harmonised standard supporting the implementation of the AI Act in the Official Journal of the European Union.

The standard, **EN 18286:2026 - Artificial intelligence - Quality management system for EU AI Act regulatory purposes**, developed by CEN and CENELEC, provides a clear legal pathway for providers of high-risk AI systems to comply with the quality management system requirements under article 17 of the AI Act.

Its **publication through Commission Implementing Decision (EU) 2026/2274** means that providers implementing this standard will benefit from a presumption of conformity with the Article 17 requirements covered by the standard.



Article 55(1) AIA

1. In addition to the obligations listed in Articles 53 and 54, providers of general-purpose AI models with systemic risk shall:

(a) perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks;

(b) assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk;

Article 56 AIA

1. The AI Office shall encourage and facilitate the drawing up of codes of practice at Union level in order to contribute to the proper application of this Regulation, taking into account international approaches.

2. The AI Office and the Board shall aim to ensure that the codes of practice cover at least the obligations provided for in **Articles 53** and **55**, including the following issues:

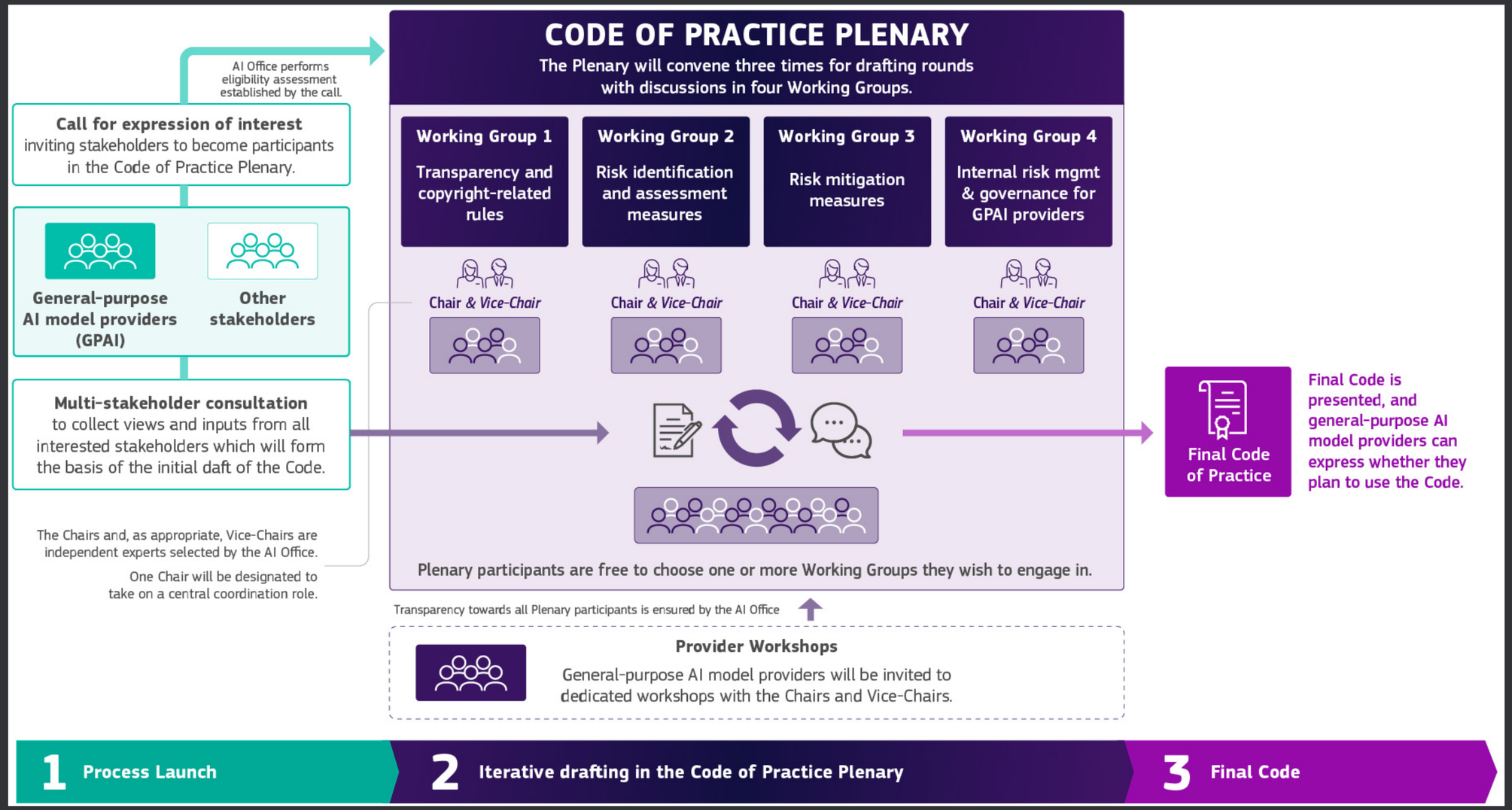
(a) the means to ensure that the information referred to in **Article 53**(1), points (a) and (b), is kept up to date in light of market and technological developments;

(b) the adequate level of detail for the summary about the content used for training;

(c) the identification of the type and nature of the systemic risks at Union level, including their sources, where appropriate;

(d) the measures, procedures and modalities for the assessment and management of the systemic risks at Union level, including the documentation thereof, which shall be proportionate to the risks, take into consideration their severity and probability and take into account the specific challenges of tackling those risks in light of the possible ways in which such risks may emerge and materialise along the AI value chain.

CoP Drafting Process



CoP Chairs & Vice Chairs

Nuria Oliver *Working Group 1* *Co-Chair*	**Alexander Peukert** *Working Group 1* *Co-Chair*	**Mattias Samwald** *Working Group 2* *Chair*	**Yoshua Bengio** *Working Group 3* *Chair*	**Marietje Schaake** *Working Group 4* *Chair*
Rishi Bommasani *Working Group 1* *Vice-Chair*	**Céline Castets-Renard** *Working Group 1* *Vice-Chair*	**Marta Ziosi** *Working Group 2* *Vice-Chair*	**Daniel Privitera** *Working Group 3* *Vice-Chair*	**Anka Reuel** *Working Group 4* *Vice-Chair*
		Alexander Zacherl *Working Group 2* *Vice-Chair*	**Nitarshan Rajkumar** *Working Group 3* *Vice-Chair*	**Markus Anderljung** *Working Group 4* *Vice-Chair*

CoP Content Part 1

Commitment 1 Documentation	Measure 6.1 Security Goal
Commitment 1 Copyright policy	Measure 6.2 Appropriate security mitigations
Commitment 1 Safety and Security Framework	**Commitment 7 Safety and Security Model Report**
Measure 1.1 Creating the Framework	Measure 7.1 Model description and behavior
Measure 1.2 Implementing the Framework	Measure 7.2 Reasons for proceeding
Measure 1.3 Updating the Framework	Measure 7.3 Documentation of systemic risks
Measure 1.4 Framework notifications	Measure 7.4 External reports
Commitment 2 Systemic risk identification	Measure 7.5 Material changes to the system
Measure 2.1 Systemic risk identification process	Measure 7.6 Model Report updates
Measure 2.2 Systemic risk scenarios	Measure 7.7 Model Report notifications
Commitment 3 Systemic risk analysis	**Commitment 8 Systemic risk responsibility and culture**
Measure 3.1 Model-independent information	Measure 8.1 Definition of clear responsibilities
Measure 3.2 Model evaluations	Measure 8.2 Allocation of appropriate resources
Measure 3.3 Systemic risk modelling	Measure 8.3 Promotion of a healthy risk culture
Measure 3.4 Systemic risk estimation	**Commitment 9 Serious incident reporting**
Measure 3.5 Post-market monitoring	Measure 9.1 Methods for serious incident reporting
Commitment 4 Systemic risk acceptance determination	Measure 9.2 Relevant information for serious incidents
Measure 4.1 Systemic risk acceptance criteria	Measure 9.3 Reporting timelines
Measure 4.2 Proceeding or not proceeding decisions	Measure 9.4 Retention period
Commitment 5 Safety mitigations	**Commitment 10 Additional documentation & transparency**
Measure 5.1 Appropriate safety mitigations	Measure 10.1 Documentation
Commitment 6 Security mitigations	Measure 10.2 Public transparency

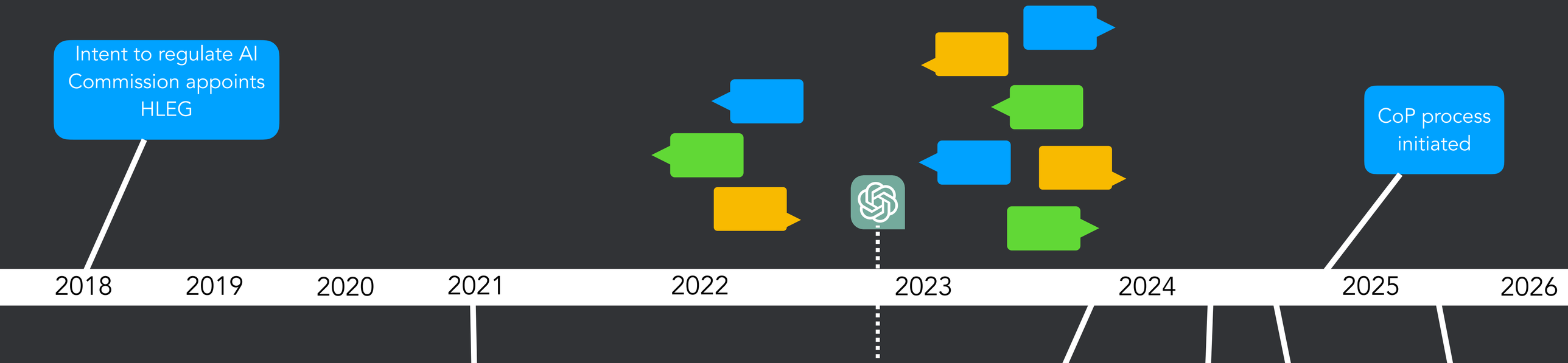
CoP Content Part 2

A 1 Systemic risks and other considerations
A 1.1 Types of risks
A 1.2 Nature of systemic risks
A 1.3 Sources of systemic risks
A 1.4 Specified systemic risks
A 2 Similarly safe or safer models
A 2.1 Safe reference models
A 2.2 Similarly safe or safer models
A 3 Model evaluations
A 3.1 Rigorous model evaluations
A 3.2 Model elicitation
A 3.3 Assessing the effectiveness of mitigations
A 3.4 Qualified model evaluation teams and processes
A 3.5 Independent external model evaluations
A 4 Security mitigation objectives and measures
A 4.1 General security mitigations
A 4.2 Protection of unreleased model parameters
A 4.3 Hardening interface-access to unreleased models
A 4.4 Insider threats
A 4.5 Security assurance

Article 56 AIA

9. Codes of practice shall be ready at the latest by 2 May 2025. The AI Office shall take the necessary steps, including inviting providers pursuant to paragraph 7. If, by 2 August 2025, a code of practice cannot be finalised, or if the AI

EU AI Act Timeline



POLITICO

rench political crisis | EU-US relations | War in Ukraine | Newsletters | Podcasts | Poll of Polls | Policy news

NEWS > TECHNOLOGY

US Big Tech plays nice with Brussels on AI rules

The EU's AI code of conduct has received broad support — except from Meta — but that wasn't a given from the outset.

Google The Keyword | Home | Product news | Company news

By Kent Walker | President of Global Affairs, Google & Alphabet

We will sign the EU AI Code of Practice.

Meta Refuses to Sign EU AI Code of Practice as Rivals Commit —Microsoft, OpenAI, Anthropic On Board

Ashley King July 21, 2025

The New York Times

Artificial Intelligence › | California's A.I. Law | Chatbots for Financial Advice | A.I. Influencers | 'Open-Source' OpenAI

European Union Unveils Rules for Powerful A.I. Systems

Makers of the most advanced artificial intelligence systems face obligations for transparency, copyright protection and public safety. The rules are not enforceable until next year.

EURACTIV | Politics | Economy | Defence | Ukraine | Rapporteur | Newsletters | Videos



POLITICS
JACOB WULFF WOLD

The Brief – EU AI code passes the Meta stress test, now it's time for hardball

Any AI rulebook Meta would voluntarily follow would be a cop-out.

External Evals ✅

Appendix 3.5 Independent external model evaluations

In addition to internal model evaluations, Signatories will ensure that adequately qualified independent external evaluators conduct model evaluations pursuant to this Appendix 3, with regards to the systemic risk, unless:

(1) the model is a similarly safe or safer model pursuant to Appendix 2.2; or

(2) Signatories fail to appoint adequately qualified independent external evaluators, despite using early search efforts (such as through a public call open for 20 business days) and promptly notifying identified evaluators, in which case Signatories will take into account the potential additional uncertainty arising from the absence of independent external evaluations (pursuant to this Appendix 3.5) when determining whether the systemic risks stemming from the model are acceptable (pursuant to Commitment 4).

Adequate qualification of independent external evaluators requires:

(1) having significant domain expertise for the systemic risk and being technically skilled and experienced in conducting model evaluations;

(2) having appropriate internal and external information security protocols in place; and

(3) having agreed to protect commercially confidential information, if they need access to such information.

Signatories will provide independent external evaluators with adequate access, information, time, and other resources (pursuant to Appendix 3.4), without prejudice to Appendix 4.4, point (1). Signatories will not undermine the integrity of external model evaluations by storing and/or analysing inputs and/or outputs from test runs without express permission from the evaluators.

The Devil Is In The Detail: EU AI Act Recital 114

***The providers of general-purpose AI models presenting systemic risks should be subject**, in addition to the obligations provided for providers of general-purpose AI models, **to obligations aimed at identifying and mitigating those risks** […]. To achieve those objectives, this Regulation should require providers to perform the necessary model evaluations, in particular prior to its first placing on the market, including conducting and documenting adversarial testing of models, **also, as appropriate, through internal or independent external testing**.*

More Context on Evals ✅

(e) all results of model evaluations relevant to understanding the systemic risks stemming from the model and descriptions of: (i) how the evaluations were conducted; (ii) the tests and tasks involved in the model evaluations; (iii) how the model evaluations were scored; (iv) how the model was elicited (pursuant to Appendix 3.2); (v) how the scores compare to human baselines (where applicable), across the model versions, and across the evaluation settings;

(f) at least five, random samples of inputs and outputs from each relevant model evaluation, such as completions, generations, and/or trajectories, to facilitate independent interpretation of the model evaluation results and understanding of the systemic risks stemming from the model. If particular trajectories materially inform the understanding of a systemic risk, such trajectories will also be provided. Further, Signatories will provide a sufficiently large number of random samples of inputs and outputs from a relevant model evaluation if subsequently asked by the AI Office;

Serious Incident Reporting ✅

Commitment 9 Serious incident reporting

LEGAL TEXT: Article 55(1), and recitals 114 and 115 AI Act

Signatories commit to implementing appropriate processes and measures for keeping track of, documenting, and reporting to the AI Office and, as applicable, to national competent authorities, without undue delay relevant information about serious incidents along the entire model lifecycle and possible corrective measures to address them, as specified in the Measures of this Commitment. Further, Signatories commit to providing resourcing of such processes and measures appropriate for the severity of the serious incident and the degree of involvement of their model.

What risks need to be assessed 🟡

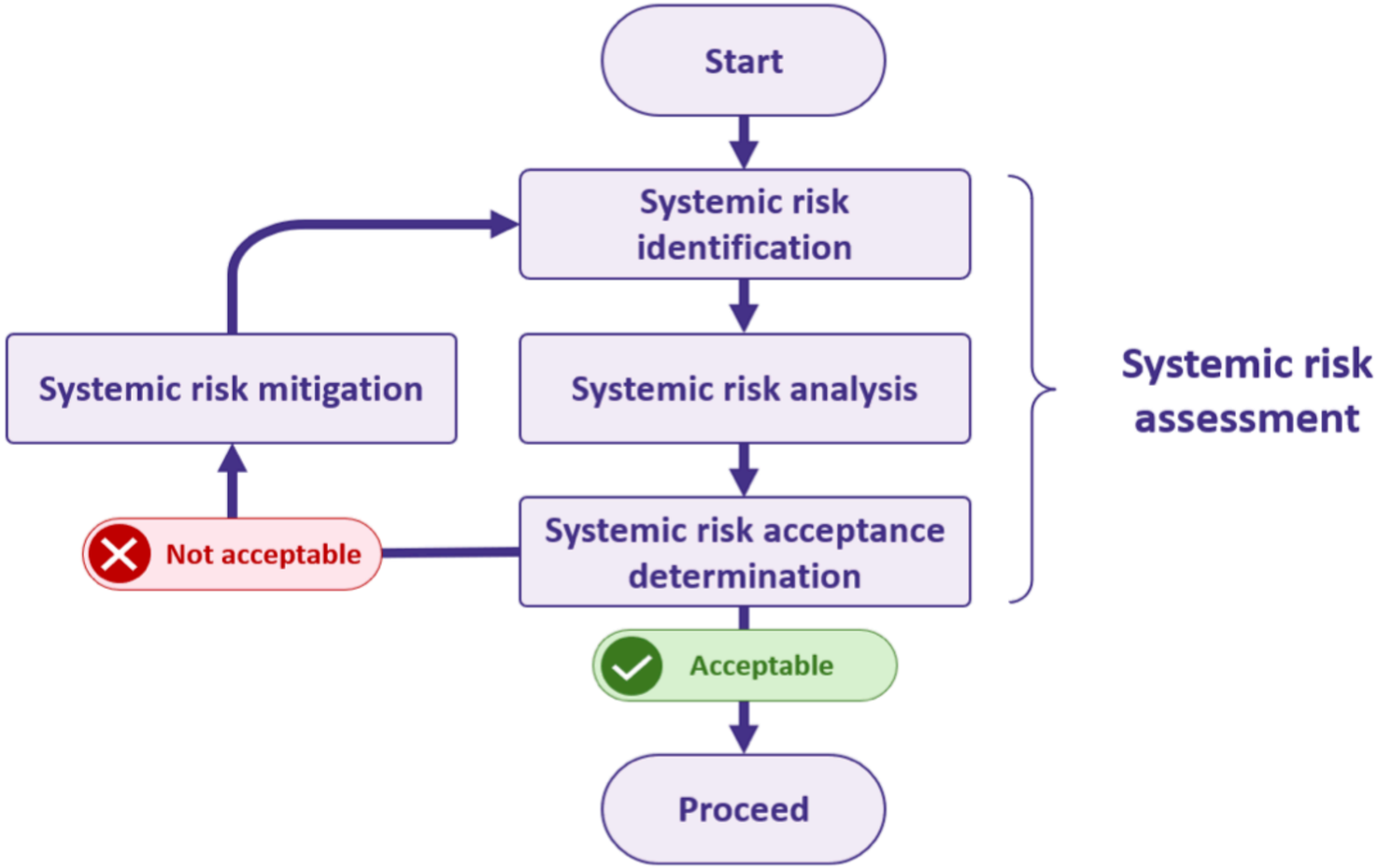


Figure 3. Full systemic risk assessment and mitigation process.
The text of the Commitments and Measures takes precedence.

Appendix 1.4 Specified systemic risks

Based on the types of risks in Appendix 1.1, considering the nature of systemic risks in Appendix 1.2 and the sources of systemic risks in Appendix 1.3, and taking into account international approaches pursuant to Article 56(1) and recital 110 AI Act, the following are treated as specified systemic risks for the purpose of systemic risk identification in Measure 2.1, point (2):

(1) **Chemical, biological, radiological and nuclear:** Risks from enabling chemical, biological, radiological, and nuclear (CBRN) attacks or accidents. This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved, in the design, development, acquisition, release, distribution, and use of related weapons or materials.

(2) **Loss of control:** Risks from humans losing the ability to reliably direct, modify, or shut down a model. Such risks may emerge from misalignment with human intent or values, self-reasoning, self-replication, self-improvement, deception, resistance to goal modification, power-seeking behaviour, or autonomously creating or improving AI models or AI systems.

(3) **Cyber offence:** Risks from enabling large-scale sophisticated cyber-attacks, including on critical systems (e.g. critical infrastructure). This includes significantly lowering the barriers to entry for malicious actors, or significantly increasing the potential impact achieved in offensive cyber operations, e.g. through automated vulnerability discovery, exploit generation, operational use, and attack scaling.

(4) **Harmful manipulation:** Risks from enabling the strategic distortion of human behaviour or beliefs by targeting large populations or high-stakes decision-makers through persuasion, deception, or personalised targeting. This includes significantly enhancing capabilities for persuasion, deception, and personalised targeting, particularly through multi-turn interactions and where individuals are unaware of or cannot reasonably detect such influence. Such capabilities could undermine democratic processes and fundamental rights, including exploitation based on protected characteristics.

Appendix 1.3.1 Model capabilities

Model capabilities include:

(1) offensive cyber capabilities;

(2) Chemical, Biological, Radiological, and Nuclear (CBRN) capabilities, and other such weapon acquisition or proliferation capabilities;

(3) capabilities that could cause the persistent and serious infringement of fundamental rights;

(4) capabilities to manipulate, persuade, or deceive;

(5) capabilities to operate autonomously;

(6) capabilities to adaptively learn new tasks;

(7) capabilities of long-horizon planning, forecasting, or strategising;

(8) capabilities of self-reasoning (e.g. a model's ability to reason about itself, its implementation, or environment, its ability to know if it is being evaluated);

(9) capabilities to evade human oversight;

(10) capabilities to self-replicate, self-improve, or modify its own implementation environment;

(11) capabilities to automate AI research and development;

(12) capabilities to process multiple modalities (e.g. text, images, audio, video, and further modalities);

(13) capabilities to use tools, including "computer use" (e.g. interacting with hardware or software that is not part of the model itself, application interfaces, and user interfaces); and

(14) capabilities to control physical systems.

Appendix 1.3.2 Model propensities

Model propensities, which encompass inclinations or tendencies of a model to exhibit some behaviours or patterns, include:

(1) misalignment with human intent;

(2) misalignment with human values (e.g. disregard for fundamental rights);

(3) tendency to deploy capabilities in harmful ways (e.g. to manipulate or deceive);

(4) tendency to "hallucinate", to produce misinformation, or to obscure sources of information;

(5) discriminatory bias;

(6) lack of performance reliability;

(7) lawlessness, i.e. acting without reasonable regard to legal duties that would be imposed on similarly situated persons, or without reasonable regard to the legally protected interests of affected persons;

(8) "goal-pursuing", harmful resistance to goal modification, or "power-seeking";

(9) "colluding" with other AI models/systems; and

(10) mis-coordination or conflict with other AI models/systems.

Public Transparency ❌

Measure 10.2 Public transparency

If and insofar as necessary to assess and/or mitigate systemic risks, Signatories will publish (e.g. via their websites) a summarised version of their Framework and Model Report(s), and updates thereof (pursuant to Commitments 1 and 7), with removals to not undermine the effectiveness of safety and/or security mitigations and to protect sensitive commercial information. For Model Reports, such publication will include high-level descriptions of the systemic risk assessment results and the safety and security mitigations implemented. For Frameworks, such publication is not necessary if all of the Signatory's models are similarly safe or safer models pursuant to Appendix 2.2. For Model Reports, such publication is not necessary if the model is a similarly safe or safer model pursuant to Appendix 2.2.

Risks to assess and mitigate ❌

LEGAL TEXT

Article 3(64) AI Act: 'high-impact capabilities' means capabilities that match or exceed the capabilities recorded in the most advanced general-purpose AI models;

Article 3(65) AI Act: 'systemic risk' means a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain;

ADDITIONAL LEGAL TEXT: Recital 110 AI Act

Appendix 1.2.1 Essential characteristics

(1) The risk **is specific to high-impact capabilities** pursuant to Article 3(65) and Article 3(64) AI Act.

(2) The risk **has significant impact on the Union market** pursuant to Article 3(65) AI Act.

(3) Said impact **can be propagated at scale across the value chain** pursuant to Article 3(65) AI Act.

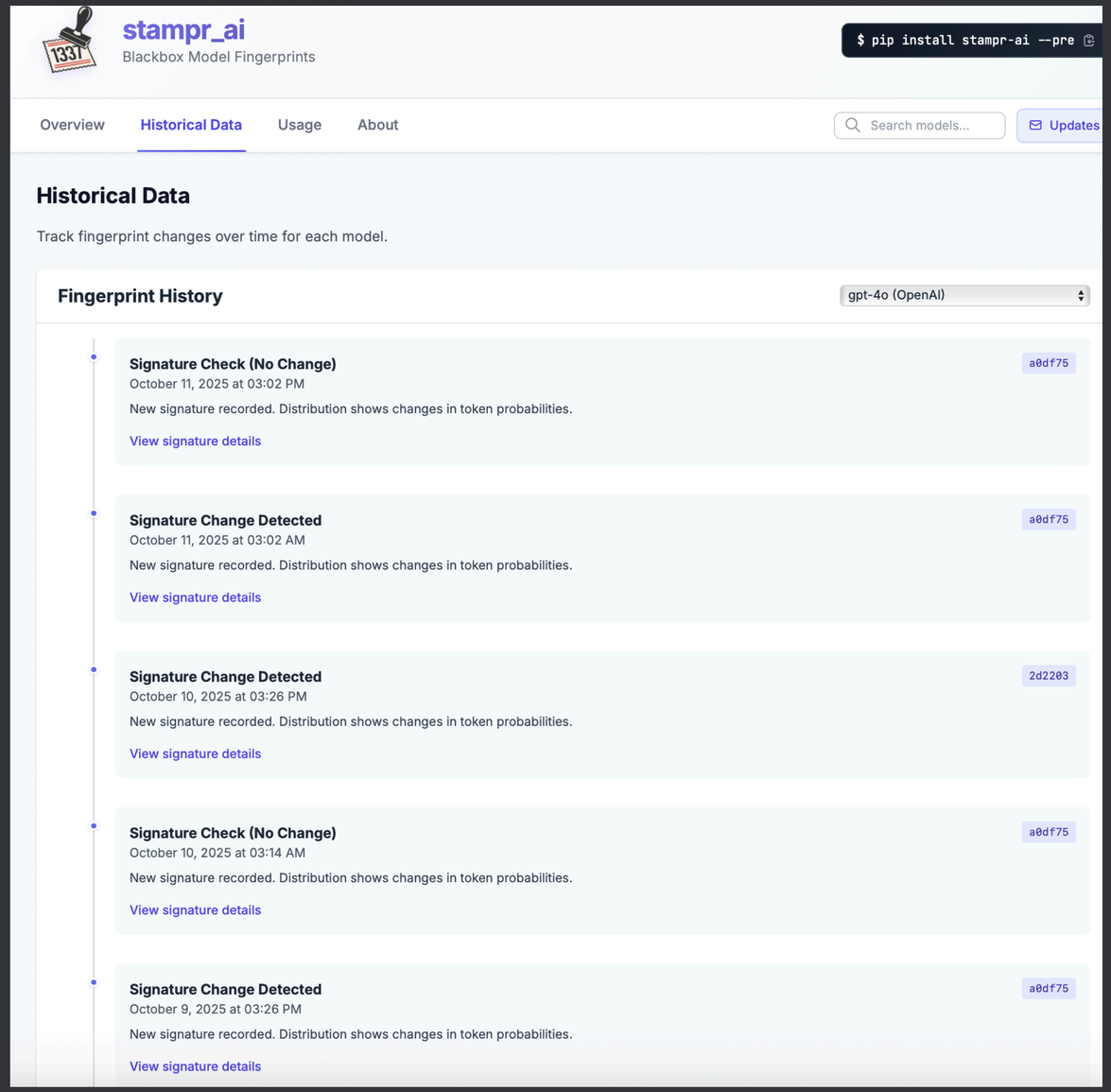
Model Versioning ❌

To facilitate post-market monitoring, Signatories will provide an adequate number of independent external evaluators with adequate free access to:

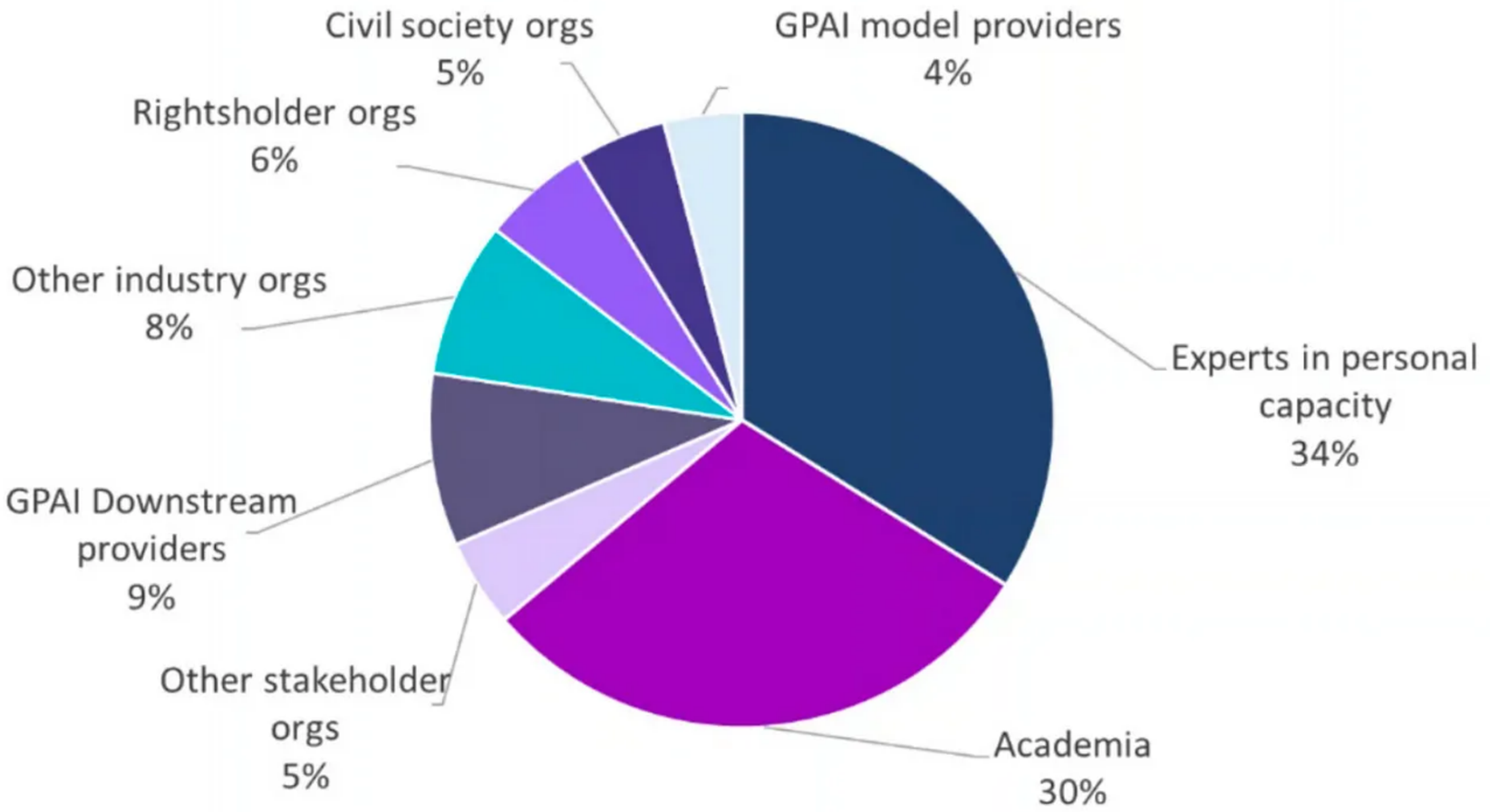
(1) the model's most capable model version(s) with regard to the systemic risk that is made available on the market;

(2) the chains-of-thought of the model version(s) in point (1), if available; and

(3) the model version(s) corresponding to the model version(s) in point (1) with the fewest safety mitigations implemented with regard to the systemic risk (such as the helpful-only model version, if it exists) and, as available, its chains-of-thought;



Privileging Model Providers ❌



Observers

International organisations:
AI Safety Institutes, OECD, UNESCO, ITU

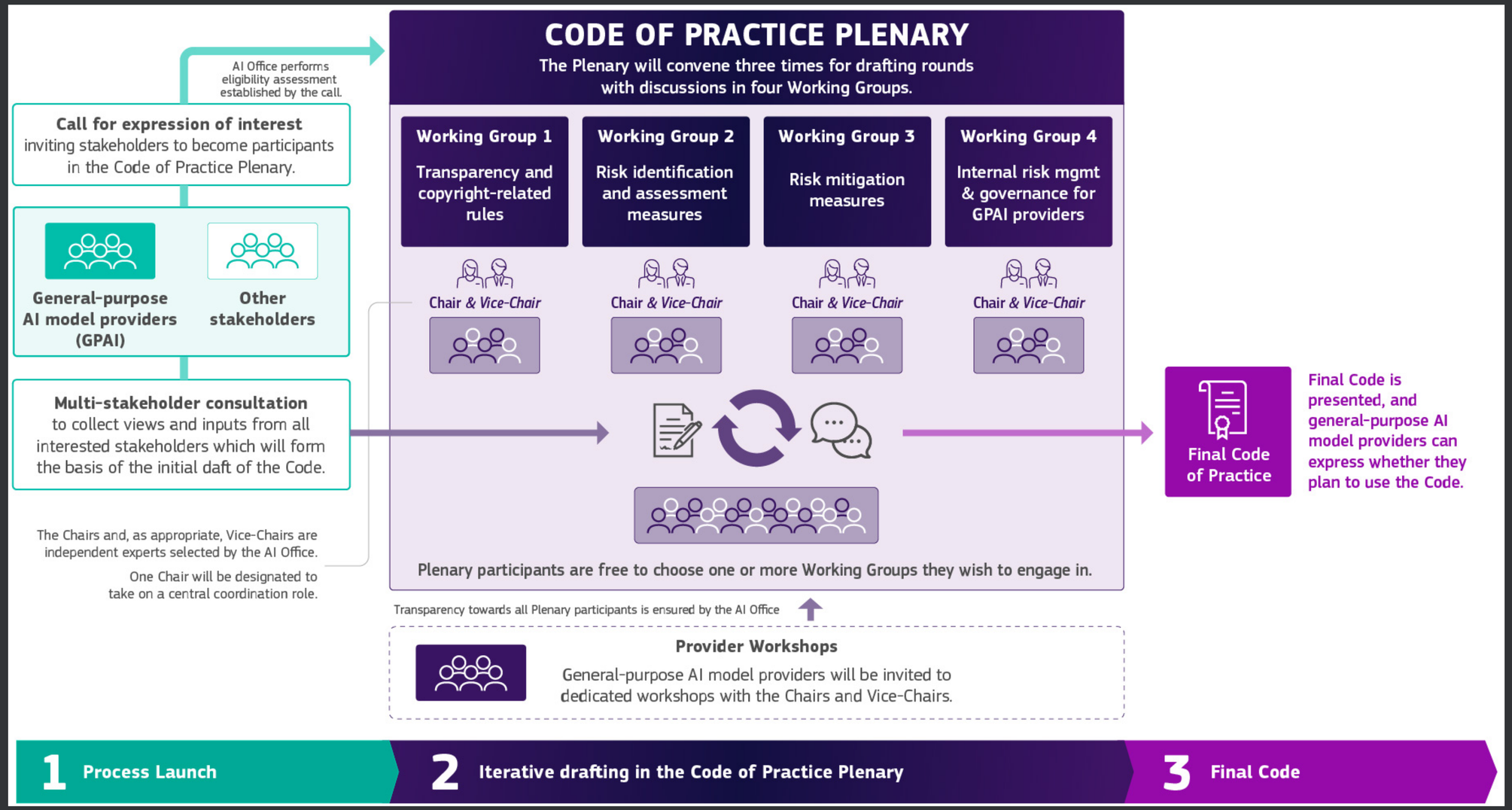
European organisations:
EUIPO, CoE, Europol, JRC, European Parliament representative

27 Member State representatives from the AI Board

Invited by AI Office – Equal status as participants

EUROPEAN ARTIFICIAL INTELLIGENCE OFFICE

CoP Drafting Process

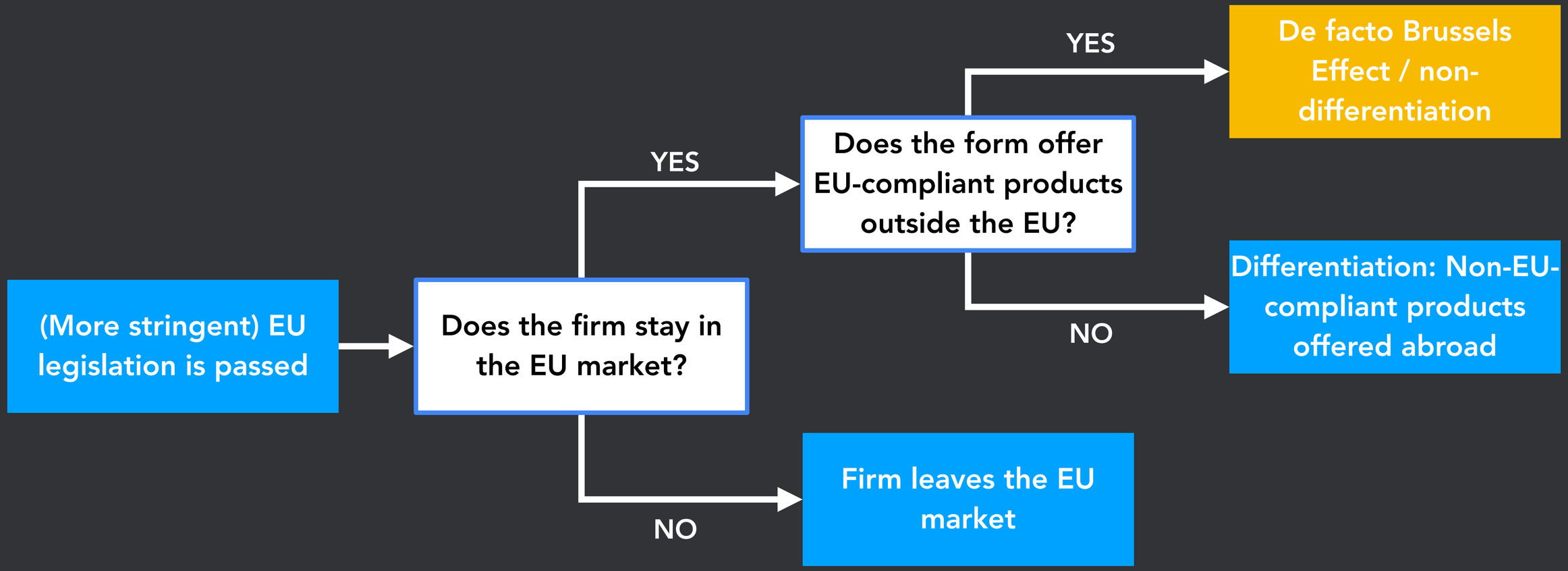


Excursion: The Brussels Effect

De Facto Brussels Effect

→ incentivising changes in products offered in non-EU countries

Excursion: The Brussels Effect



Source: Siegmann & Anderlsjung (2022)

Excursion: The Brussels Effect

De Facto Brussels Effect

→ incentivising changes in products offered in non-EU countries

De Jure Brussels Effect

→ influencing regulation adopted by other jurisdictions

The Practical Realities of AI Governance (Pt 1)

- Trust is your #1 currency; the earlier you can build i
  horizontal vs. vertical networking)
- Create win-win situations
- Treat everyone as a stakeholder
- There’s a lot of uncertainty you need to be able to
  decision-making. You will make mistakes. Everyone
- Things will heat up; don’t be angry if people play t



The Practical Realities of AI Governance (Pt 2)

- At best, everyone will be equally unhappy
- Make sure journalists know your name and how to connect so that you can provide your side of the story
- There’s communicated red lines and there’s actual red lines
- Who are you negotiating with? Same people who make the final decision?
- Create (perceived) wins → put provisions in that you’re willing to cut or sell as win; make content look shorter

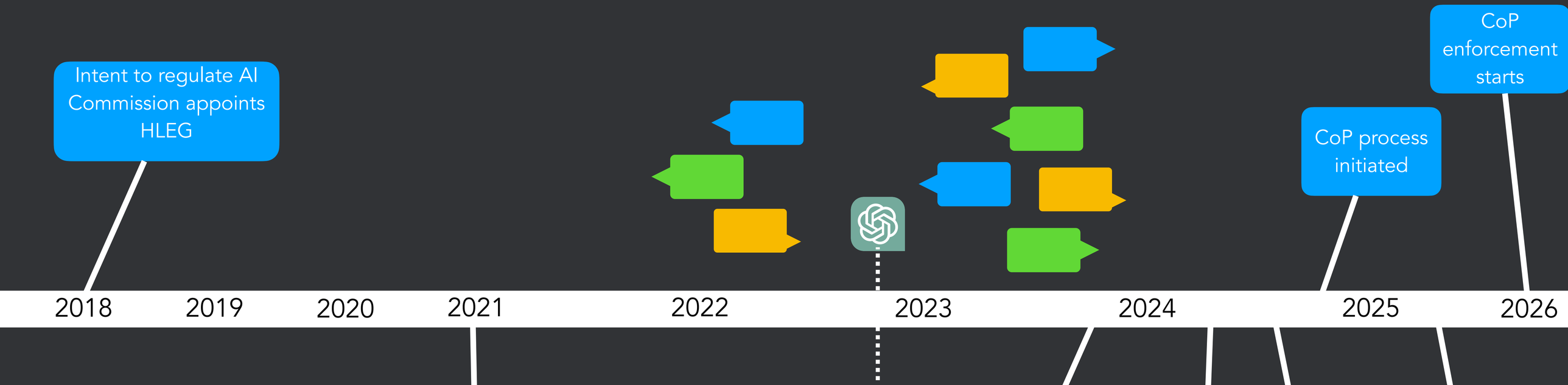
The Practical Realities of AI Governance (Pt 3)

- Meta publicly said after third drafting round: “We’re not signing.” → made negotiations for working-level teams much harder in the arguably most important drafting round
- Optics matter (e.g., length) → support internal stakeholder management for trusted contacts
- Future-proofing the Code as necessity due to a lack of implemented updating mechanisms → adaptive governance as necessity for AI

Okay, so we now have the AIA and the CoP – so we're good, right?

About Enforcement & Political Will

EU AI Act Timeline



About Enforcement & Political Will



BBC

Home News Sport Business Innovation Culture Arts Travel Earth Audio Video Live

Trump says Apple boss called him to complain about EU

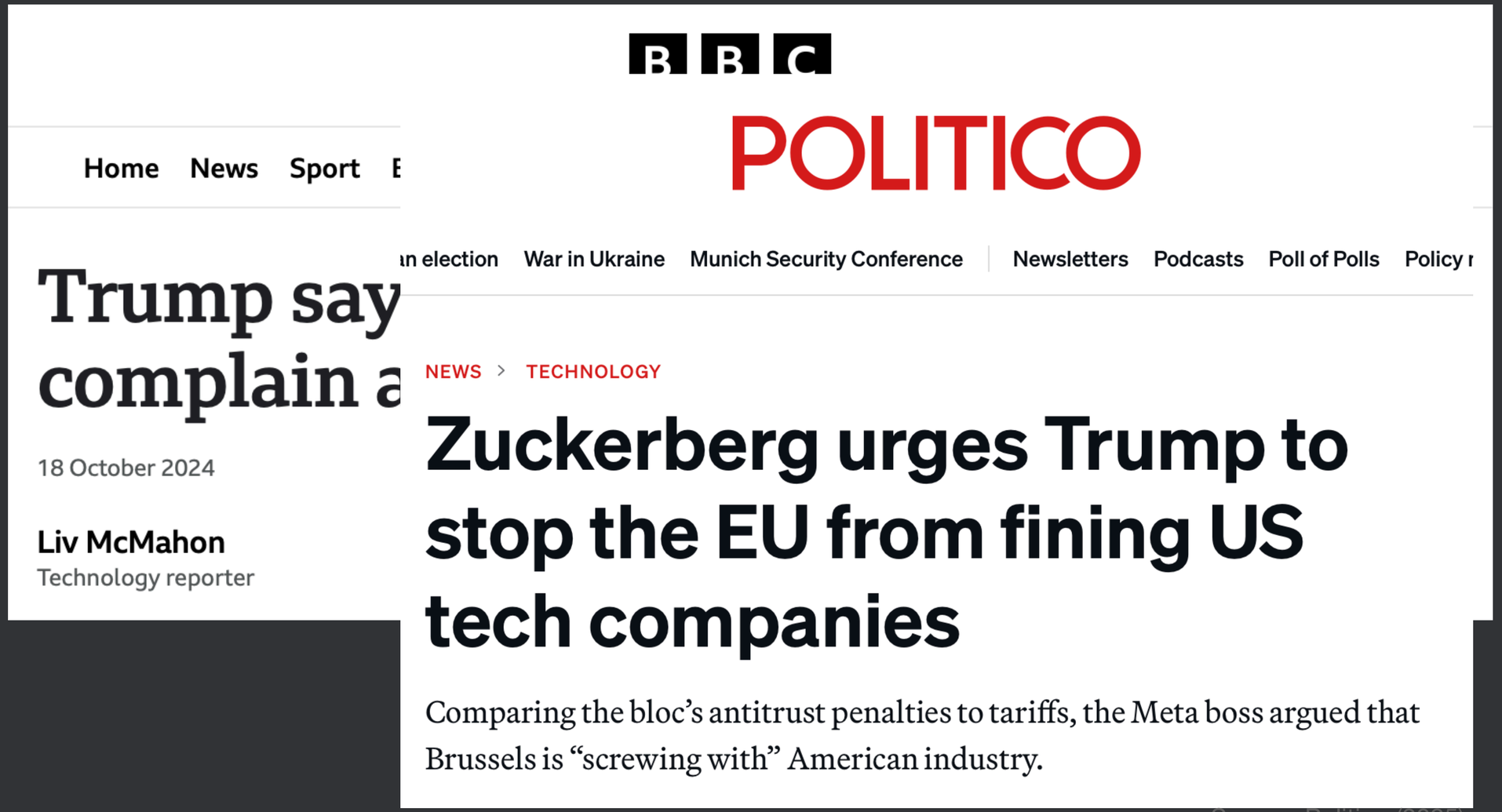
18 October 2024

Share Save

Liv McMahon
Technology reporter

Source: BBC (2024)

About Enforcement & Political Will



Source: Politico (2025)

About Enforcement & Political Will

Home News Sport E

Trump say
complain a

18 October 2024

Liv McMahon
Technology reporter

FINANCIAL TIMES

ME WORLD US COMPANIES TECH MARKETS CLIMATE OPINION LEX WORK & CAREERS LIFE & ARTS HTSI



NEWS

Z
st
tech companies

Comparing the bloc’s antitrust penalties to tariffs, the Meta boss argued that Brussels is “screwing with” American industry.

Source: Financial Times (2025)

About Enforcement & Political Will



Source: Financial Times (2025)

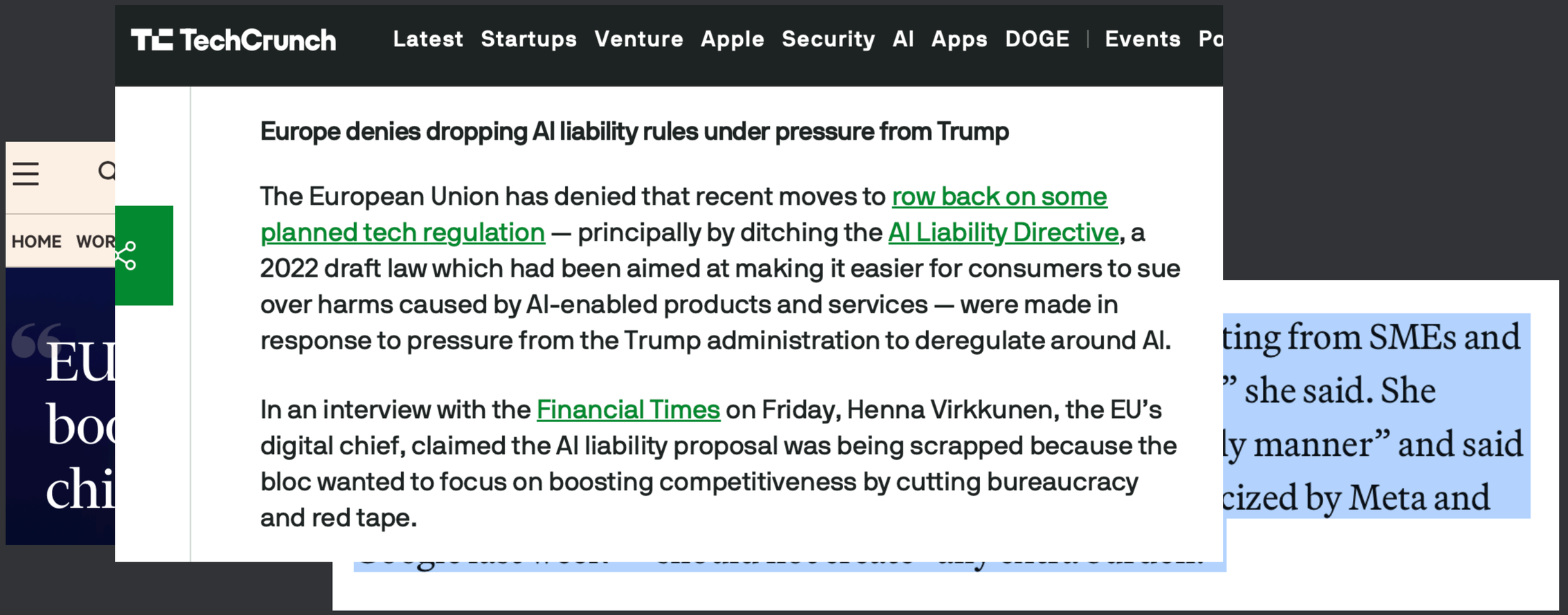
About Enforcement & Political Will



“I’m also taking this criticism very seriously that we’re getting from SMEs and industry that we have too much bureaucracy and red tape,” she said. She pledges to implement the AI Act “in an innovation-friendly manner” and said that the EU’s voluntary code of practice — which was criticized by Meta and Google last week — should not create “any extra burden.”

Source: Politico (2025)

About Enforcement & Political Will



Source: TechCrunch (2025)

Digital Omnibus (Regulation 2026/1744)

- Defers the stand-alone high-risk AI obligations from Aug 2 2026 to Dec 2 2027 (likely because of pressure & standards not ready)
- New prohibition for AI systems used to generate child sexual abuse material (CSAM) or non-consensual sexual or intimate content → responsibility to implement safeguards such as refusal training, output controls and content filtering
- AI built into machinery, such as industrial robots, no longer has to meet the AI Act's high-risk requirements directly
    - Manufacturers face less overlapping regulation but creates precedent for other product AI

By Luca Bertuzzi

Published on 14 October 2026 - 10:35 GMT+2 • Updated 11:38

Share Comments Add Euronews on Google

The EU is investigating OpenAI and Anthropic over AI agents that hacked outside organisations, raising questions about whether the AI Act covers rogue models escaping the testing environment.

The European Commission is moving to enforce its new rules on leading AI labs after a string of incidents, EU digital chief Henna Virkkunen said. But applying the rulebook is far from straightforward, and these early cases will set important precedents.

Art. 2(8)

This Regulation does not apply to any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service. Such activities shall be conducted in accordance with applicable Union law. Testing in real world conditions shall not be covered by that exclusion.

- AI Act does *not* cover internal deployments
  - BUT models that were on the EU market at the time were involved in incidents
  - BUT models were arguably tested in the real world (Art. 2(8) doesn’t apply)
  - BUT Hugging Face hosts AI models and datasets that developers use globally, i.e., arguably the OpenAI incident impacted EU customers
- Model providers can be fined up to 3% of worldwide annual turnover or €15 million, whichever is higher

Not perfect but one of the most serious attempts at governing AI we have seen so far
(and it gives us an idea about the practical challenges)

Five ways to govern catastrophic risk



1

Precautionary principle

Act before certainty



2

Licensing

Prove safety before market



3

Self-governance

Professional norms and responsibilities



4

Nonproliferation

Control the critical inputs



5

Global treaties

Bind states, verify compliance

1

The Precautionary Principle

When should uncertainty justify action?

What is the precautionary principle?

When an activity threatens serious or irreversible harm, act to prevent it, even before the science is certain

Origin: West Germany's Vorsorgeprinzip ("foresight principle"), 1970s environmental policy



Threat

Possible harm is severe or irreversible



Uncertainty

Cause and effect not fully established



Action

Protective measures taken anyway

Two versions: weak and strong

WEAK

Uncertainty is no excuse for inaction

- Permits action; does not require it
- Burden of proof stays with regulators
- Apply cost-benefit analysis

Rio Declaration, Principle 15 (1992)

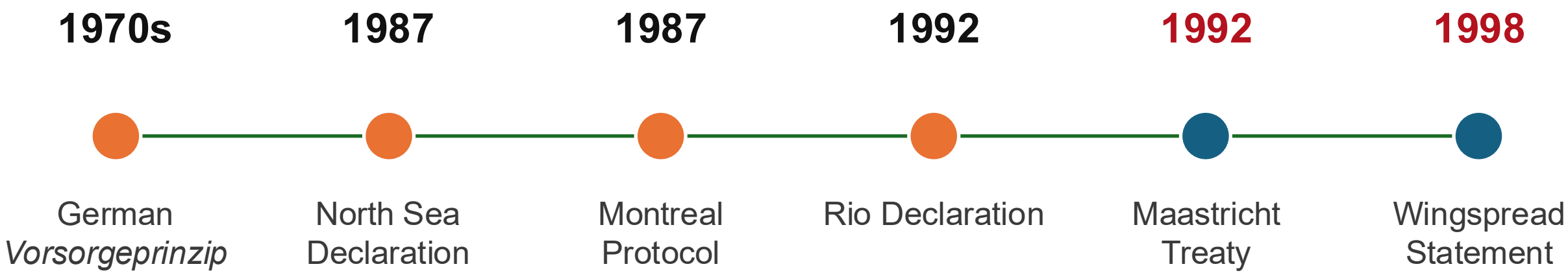
STRONG

Unproven safety is a reason to restrict

- Requires precaution under uncertainty
- Burden of proof shifts to proponents
- Can justify bans or moratoria

Wingspread Statement (1998)

Born of environmental harm



From national planning idea to binding treaty law in under 25 years

Maastricht (1992): precaution as law



Article 130r(2)

EU environmental policy “shall be based on the precautionary principle”

Binding status

First treaty-level legal basis in a major polity; now Article 191 TFEU

Beyond environment

EU courts extend it to food and health, e.g. the BSE (mad cow) beef ban

Disciplined use

2000 Commission guidance: proportionate, non-discriminatory, consistent, cost-aware, reviewable

Wingspread (1998): the strong version

Scientists, lawyers and activists, Racine, Wisconsin, January 1998



Act under uncertainty

Take precautions before cause and effect are proven



Shift the burden

Proponents, not the public, must show safety



Weigh alternatives

Including the option of not proceeding at all



Decide openly

Democratic, transparent, inclusive process

Sunstein: strong precaution paralyzes

Risks are on all sides. Regulation creates risks too, so a rule against risky action forbids itself.

Cass Sunstein, Laws of Fear (2005)

Why it still feels right: availability heuristic, probability neglect, loss aversion

Precaution on one side, risk on the other

Delay new drugs	→	Patients go untreated
Restrict nuclear power	→	More fossil fuel pollution
Ban DDT	→	Malaria control weakens

Sunstein’s Anti-Catastrophe Principle

Keep precaution, but reserve it for catastrophe

1

Catastrophic stakes

Harm on a massive scale, not ordinary risk

2

Deep uncertainty

Probabilities cannot be reliably assigned

3

Irreversibility

Lost options cannot be recovered later

Response: buy a margin of safety, while still weighing costs

AI: a test case for precaution

10%

Hinton's estimated chance AI causes human extinction

>10%

Anthropic's Hubinger: chance AI kills all humans this decade

1%

Bengio: even this risk should be "code red"

Catastrophic + uncertain + irreversible = squarely in anti-catastrophe territory

Leading in the Face of Uncertainty

- Yoshua Bengio:

- We should be guided by the precautionary principle:

- When potential risks and harms are catastrophic or existential, a lack of scientific certainty should not delay preventive measures.



Source: Wikipedia CC license

Bengio’s precautionary argument

Severe stakes

Risks up to human extinction

→

Uncertain odds

Probabilities debated, not settled

→

Extreme caution

Safety approach commensurate with existential stakes

Precedents he cites

Biology’s caution on gain-of-function research and mirror bacteria. Think here of Doudna!

Leading in the Face of Uncertainty

Marc Andreessen

”Our enemy is the precautionary principle”



← **Post**

Marc Andreessen @pmarca

The Precautionary Principle: The haunting fear that someone, somewhere, may be doing something innovative.

5:54 PM · Oct 27, 2023 · **120.9K** Views

Source: Wikipedia CC license

2

Licensing and Pre-Market Authorization

Should developers prove safety before release?

Precedents: drugs and nuclear power



Drugs: the FDA

Trigger: thalidomide birth defects

- 1962 Kefauver-Harris Act
- Sponsors must prove safety and efficacy
- Phased clinical trials before approval



Nuclear: the Nuclear Regulatory Commission

- 1974: NRC split from the AEC
- Separate construction and operating licenses

Shared lesson: the burden of proof sits with the producer, and the regulator must be independent of the promoter

Licensing AI: proposals and problems

Proposals on the table

Hassabis

Independent FINRA-style body to test models before release

Hinton

FDA-style approval, like pharmaceuticals

Anthropic

National testing law with power to block unsafe models

Why AI is harder

No safety test

Evaluation science is immature

Speed

Release cycles in months, not years

General purpose

One model, countless uses

Capture

Who funds the tester?

3 Self-Governance and Professional Norms

AI self-governance: the 2026 accord

4

layers of voluntary controls in the White House Accord

1,386

lab employees signed the “Pacing the Frontier” declaration

0

legal penalties for non-compliance

Strengths

Fast; draws on expert knowledge; industry-run standards body planned

Weaknesses

No mechanism for enforcement beyond imposing reputational costs

Compare biology’s record of self-restraint

EXPERT TAKE DIGITAL POLICY

Trump's AI Safety Pact Is Toothless. But There Is a Path Forward.

The White House AI accord outlines sensible safeguards, including independent evaluators and board-level oversight, but does not compel companies to adopt them. Until the cost of safety failures outweighs the payoff of racing ahead, a pledge built on "believe" and "should" won't slow anyone down.



Dario Amodei, co-founder of Anthropic, takes questions alongside Mark Zuckerberg, CEO of Meta Platforms, and U.S. President Donald Trump as Trump holds a press briefing with AI executives following a meeting on artificial intelligence outside the West Wing of the White House on September 29, 2026 in Washington, DC. Kevin Dietsch / Getty Images

Language Processing, Machine Learning

The Time Is Now to Develop Community Norms for the Release of Foundation Models

Perspectives about the benefits and risks of release vary widely. We propose setting up a review board to develop community norms and encourage coordination on release for research access.

May 17, 2022 | Percy Liang, Rishi Bommasani, Kathleen Creel, and Rob Reich

4

Nonproliferation and Supply-Side Controls

Can we control the critical inputs?

Nuclear: control the fissile material

The bottleneck is not knowledge but material: highly enriched uranium and plutonium are scarce, detectable and trackable

Outcome: about 9 nuclear-armed states, versus fears of 15–25 by the 1970s



IAEA safeguards

Material accounting and on-site inspections



Nuclear Suppliers Group

Nuclear export controls, formed after India's 1974 test



Physical protection

Securing stockpiles against theft

AI: export controls on advanced chips

Oct 2022	US restricts top AI chips to China
Oct 2023	Loopholes closed; compliant variants banned
May 2026	Overseas-subsidiary loophole closed
2026	Remote cloud access still uncontrolled

Chips are not plutonium

Leakage

Diversion via third countries; cloud rental

Substitution

Huawei and others fill the gap

Efficiency

Better algorithms need less compute

5

Global Governance and Treaties

Can rival states bind themselves, and verify it?

Two treaty models: CWC and NPT

1993 · in force 1997

Chemical Weapons Convention

- Total ban on a whole weapon class
- OPCW inspectors; challenge inspections
- All declared stockpiles destroyed by 2023
- Yet: Syria’s attacks, Novichok poisonings

1968 · in force 1970

Non-Proliferation Treaty

- A bargain, not a ban: five recognized nuclear states
- IAEA safeguards verify civilian programs
- Near-universal: 191 parties
- Yet: India, Pakistan, Israel outside; North Korea left

AI diplomacy in 2026: first steps

Sept 20	First US-China AI dialogue, New York
Sept 23	UN Security Council hears Bengio, Amodei, Altman
Sept 25	Xi and Trump agree an AI incident channel
Nov	Formal US-China AI dialogue scheduled

The missing piece: verification

Inspectors can count centrifuges and shells. No one yet knows how to verify what is trained inside a data center.

Trump: the US is "not going to be putting on brakes"

Defense in depth: no single tool

Principle	Precaution sets who bears the burden of proof
Firms	Norms and audits catch problems early
States	Licensing and chip controls enforce limits
World	Treaties stop racing and defection

Open question: can governance move as fast as the technology?