US AI Policy

Law 4052 // Computer Science 283 // Communication 252A // Political Science 145B/445B
Fall Quarter 2026



Rob Reich
Anka Reuel
Nate Persily

AI Governance in Two Sentences

"The real problem of humanity is the following: we have…

Paleolithic emotions,

medieval institutions,

and **godlike technology.**

And it is terrifically dangerous,
and it is now approaching a point of crisis overall."

E.O. Wilson, Harvard University, 2009

Shaping the scientific and technological frontier

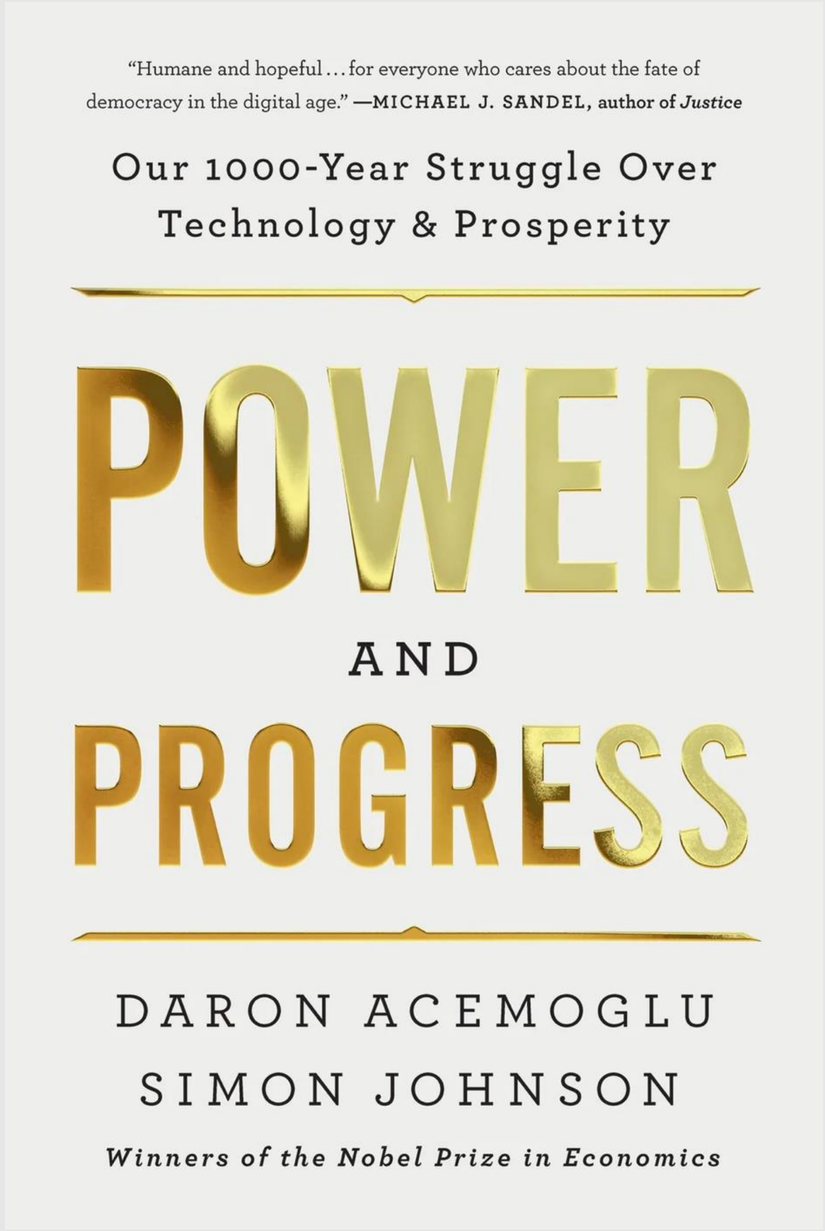
- Nothing about the path of science or technology is inevitable
- Institutional choices shape everything
- In technology, distinguish:



Our AI future is about **institutional choices** to shape and steer the technology, not predictions to react to some inevitable technological outcome

It is a **DESIGN** problem, not a **PREDICTION** problem

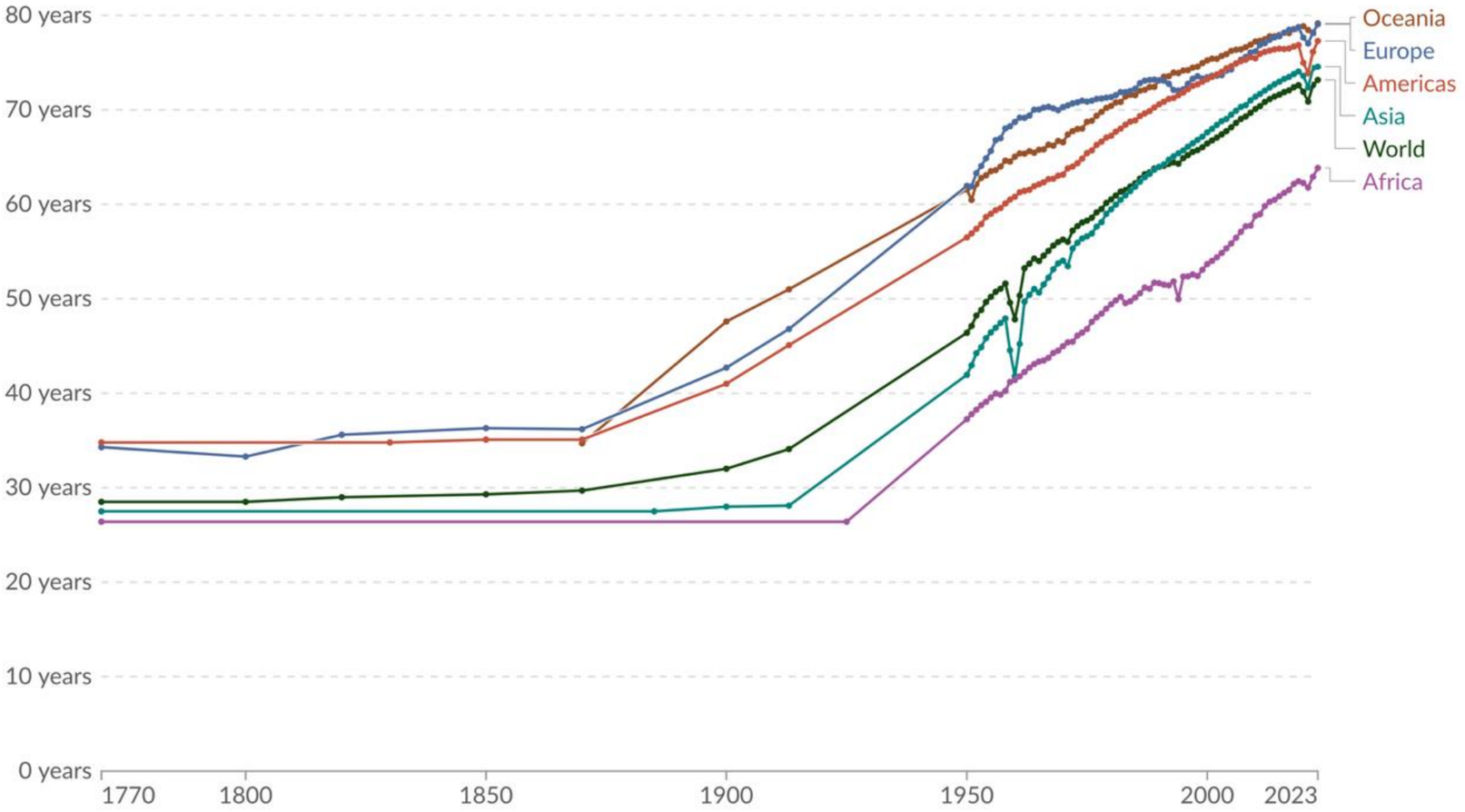
Institutions and incentives matter



Life expectancy

Our World in Data

Period life expectancy[1] is the number of years the average person born in a certain year would live if they experienced the same chances of dying at each age as people did that year.



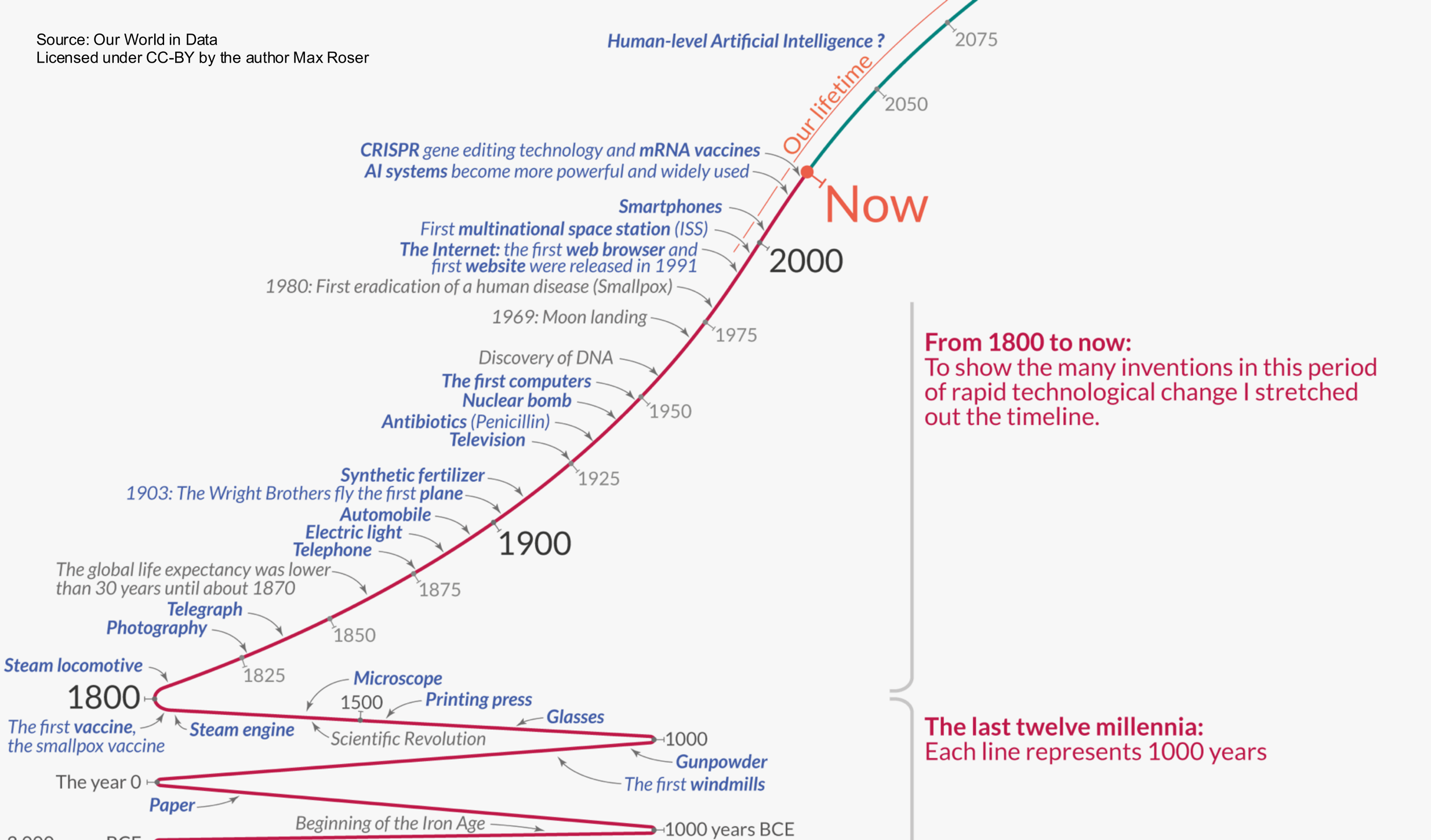
Data source: Riley (2005); Zijdeman et al. (2015); HMD (2024); UN WPP (2024)

OurWorldinData.org/life-expectancy | CC BY

1. Period life expectancy Period life expectancy is a metric that summarizes death rates across all age groups in one particular year.
For a given year, it represents the average lifespan for a hypothetical group of people, if they experienced the same age-specific death rates throughout their whole lives as the age-specific death rates seen in that particular year.
Learn more in our articles:

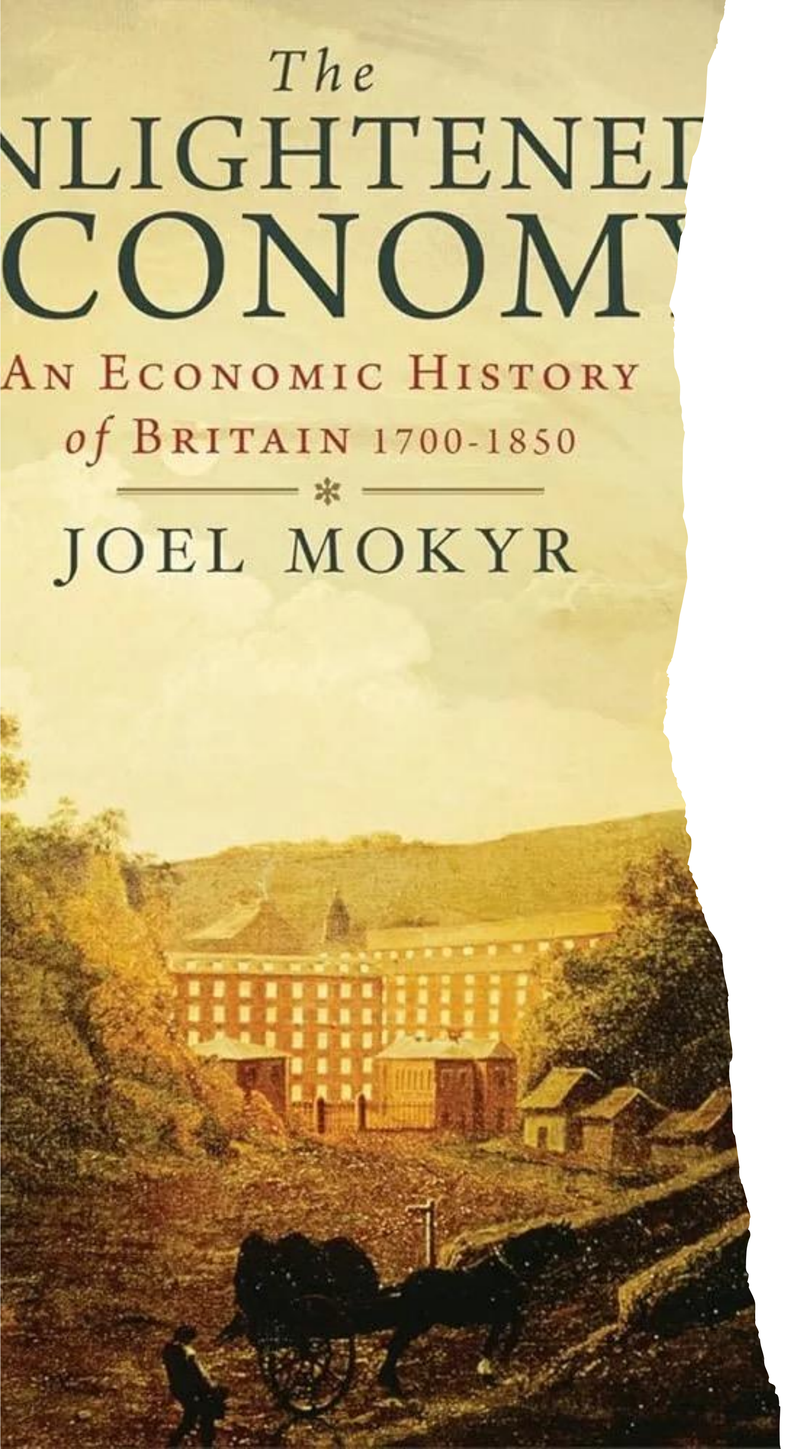
- Life expectancy – what does this actually mean?
- Period versus cohort measures: what's the difference?

Source: Our World in Data
Licensed under CC-BY by the author Max Roser



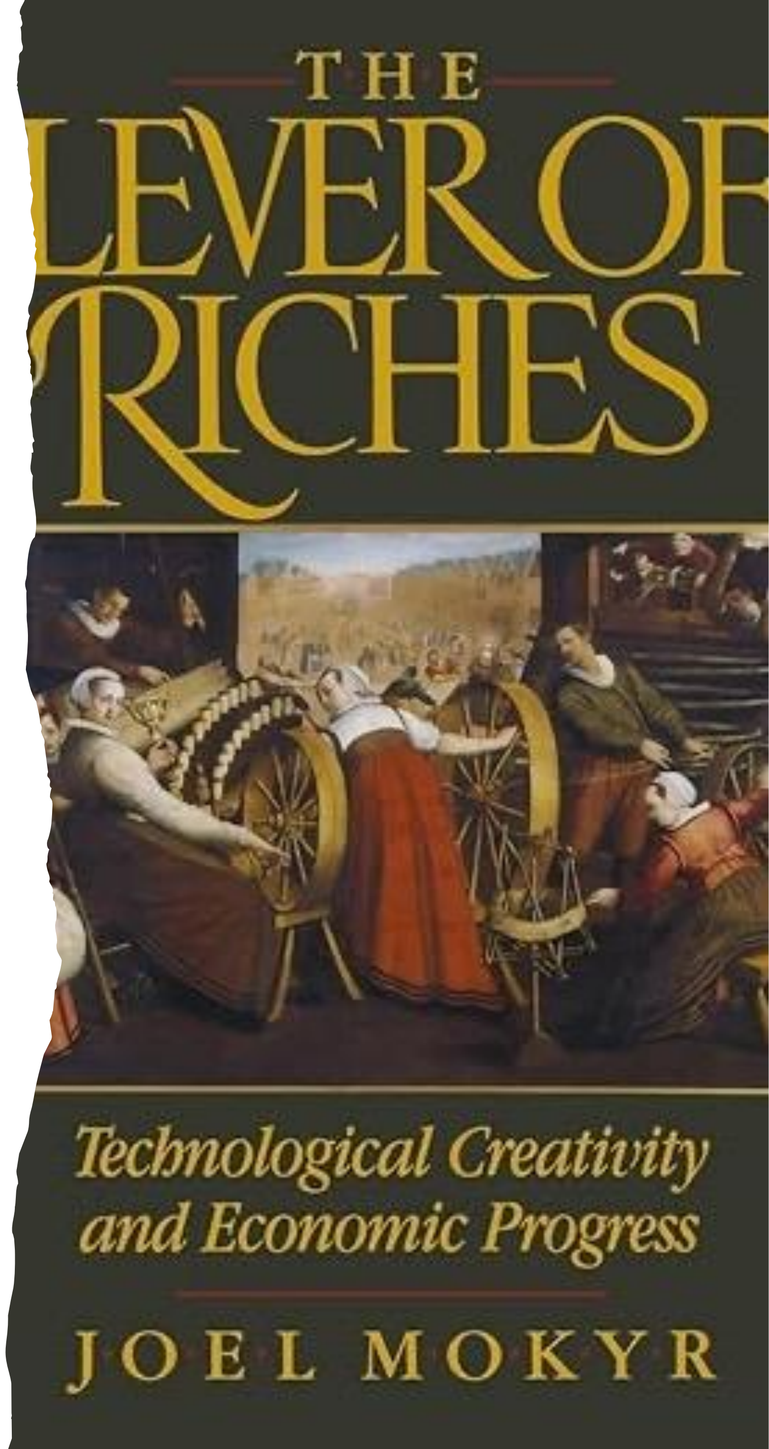
From 1800 to now:
To show the many inventions in this period of rapid technological change I stretched out the timeline.

The last twelve millennia:
Each line represents 1000 years



Joel Mokyr, Nobel Prize in Economics 2025

Economic historian who explained the role of technological innovation and the institutional arrangement of knowledge that combine to spur growth and prosperity



Regulating Frontier Science and Technology in the United States

- In the historical context of emerging technologies, innovation and market dissemination usually precedes democratic regulation

- A contrast with the European approach, more rooted in the precautionary principle

In the US, delayed regulation is a feature, not a bug

- A whistlestop tour of the US approach to regulating science and technology.
- Focus on widespread adoption and dissemination, where the benefits have been made tangible and the risks or harms are also clear.



Railroads → Telephone → Airplane → Shipping containers → **Generative AI**

Railroads

1804

The world’s first steam-powered locomotive journey

1827

First regular carrier of passengers and freight via the Baltimore and Ohio Railroad

1887

Congress passes the Interstate Commerce Act, making railroads the first industry subject to federal regulation

60 years from first passenger service to federal regulation



Telegraph and Telephone

1830s
The telegraph is invented

1851
Western Union Telegraph Company is established

1876
Alexander Graham Bell patents the telephone

1885
American Telephone and Telegraph Company (AT&T) is formed

1934
The Communications Act of 1934 establishes the FCC

58 years from Bell's patent to the FCC



Airplane

1903
Wilbur and Orville Wright invent and fly the Wright Flyer

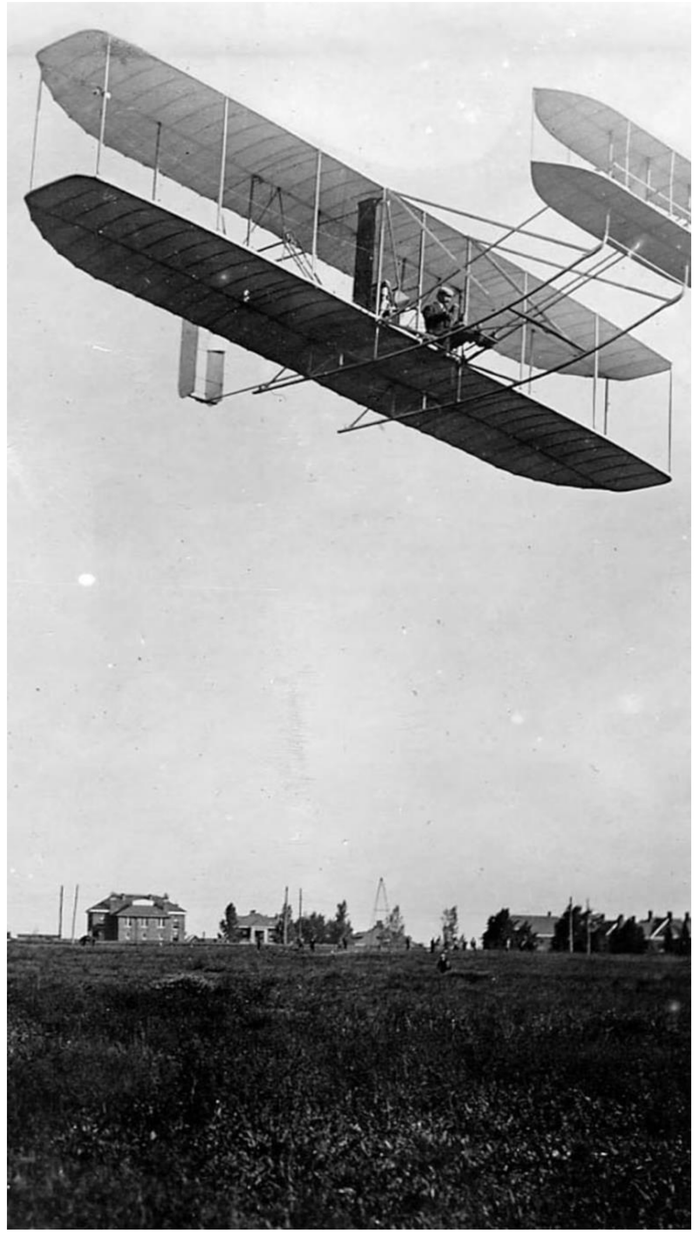
1908
First commercial flight takes off

1920s
Commercial flights become more regular

1935
Qantas opens its first international flight

1958
The Federal Aviation Administration (FAA) is established

55 years from the Wright Bros to the FAA



Headlines in History

June 30, 1956: Mid-Air Collision Over the Grand Canyon

May 28, 2026 Jenny Ashcraft

On June 30, 1956, two commercial airliners flying at 21,000 feet collided over the Grand Canyon in Arizona. The crash killed 128 people and was the deadliest aviation disaster in U.S. history at the time. The tragedy led to major changes in aviation safety and the creation of what is now the Federal Aviation Administration.

In 1956, commercial airline travel was still a novelty for most Americans. Fewer than one in three people flew that year.



Worst Commercial Tragedy in History

From Wire Dispatches

WINSLOW, Ariz. — The wreckage of one of two big airliners which disappeared Saturday on flights east from Los Angeles was spotted Saturday night in northern Arizona's Grand Canyon.

Lynn Coffin, chief ranger at Grand Canyon National Park, said the plane was identified as a TWA Super Constellation carrying 70 person. Still missing was a United Airliner carrying 58 persons.

Evansville Press: July 1, 1956

On Saturday morning, June 30, TWA Flight 2 departed Los Angeles International Airport at 9:01 am. The Lockheed Super Constellation carried 70 crew and passengers and was headed to Kansas City. Three minutes later, United Flight 718, a DC-7 headed for Chicago, also departed Los Angeles International Airport carrying 58 crew and passengers.

The TWA flight climbed to 19,000 feet, and the United flight climbed to 21,000 feet, giving both planes plenty of space between them. Still over California, the TWA pilot requested permission to climb to 21,000. His request was denied, as it would put both planes at the same altitude. When the TWA flight once again requested permission to fly above the weather, possibly to avoid thunderstorms, the request was granted. Air Traffic Control (ATC) assumed that clear skies would allow both pilots to see the other aircraft. Neither pilot was apprised of the other's position.



Shipping containers

1956

Malcolm McLean launches the first intermodal container; the shipping industry now moves cargo, not sailing ships

1960s

First purpose-built container ships

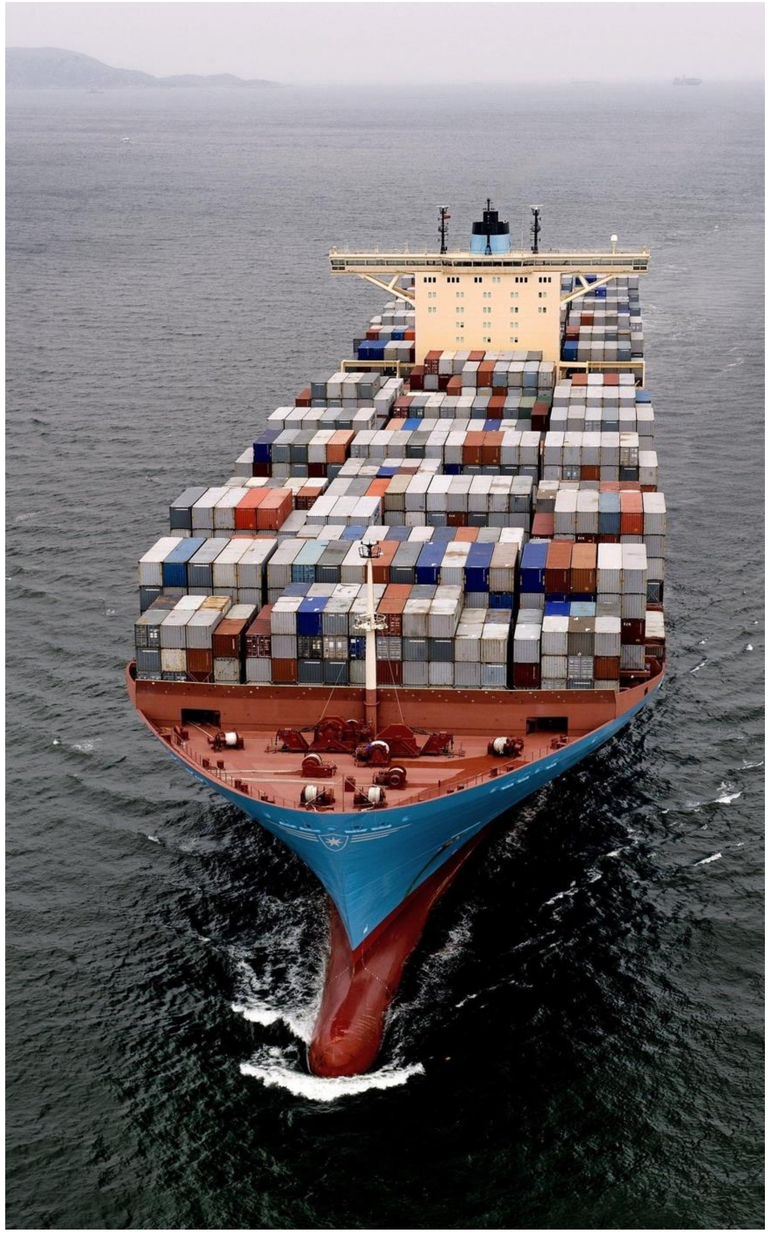
1970s

Adoption of International Organization for Standardization (ISO) container standards

1984

US passes the Shipping Act, replacing 1916 legislation

28 years from the first container to the Shipping Act



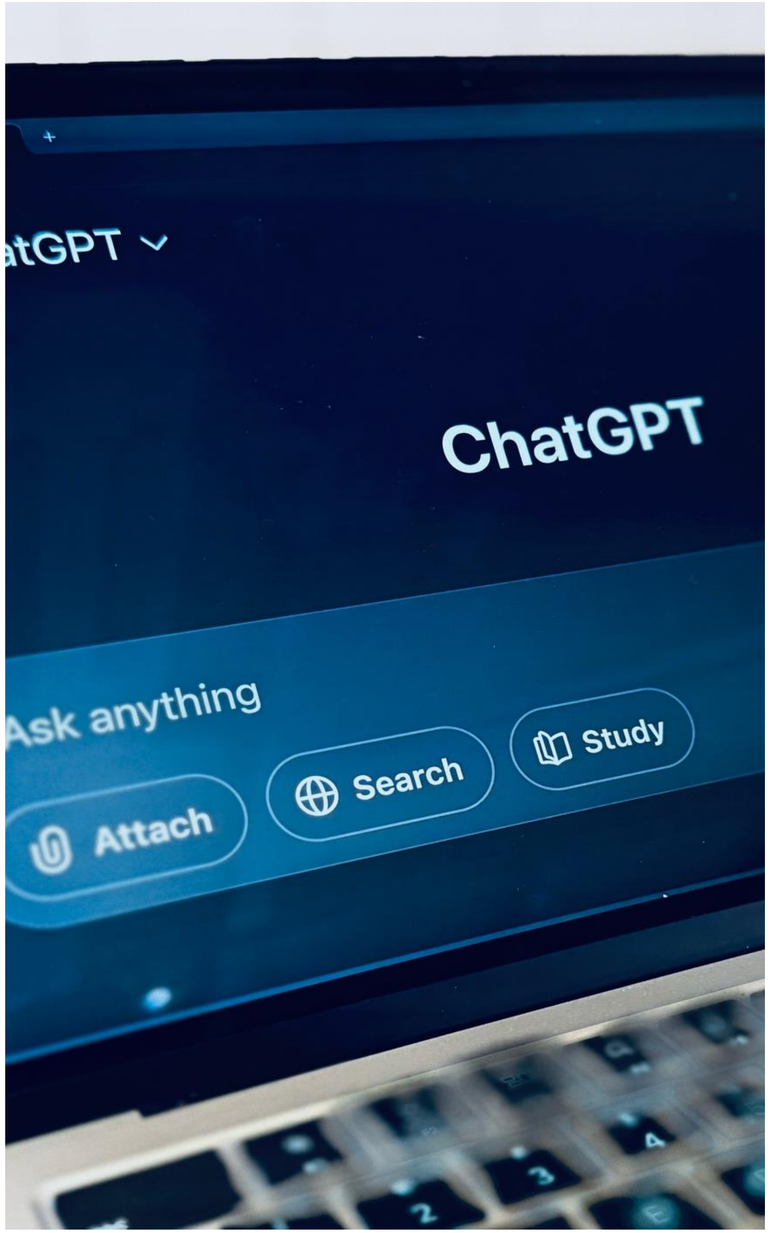
Generative AI

2022
OpenAI releases ChatGPT

2023
Claude, Gemini, LLaMa, DeepSeek

2026
1B weekly users of ChatGPT

Policy Window has opened



From innovation to dissemination in 4 years

In the US, Democracy is right on time: the policy window has opened for AI

2016	The EU General Data Protection Regulation (GDPR), the strongest privacy and security law in the world
2023	Biden White House releases Executive Order on Artificial Intelligence
2024	Creation of the AI Safety Institutes
2025	EU AI Office's General-Purpose AI Code of Practice
2025	Trump White House releases AI Action Plan, era of AI accelerationism and AI sovereignty
2026	New Trump Executive Order on AI
What's next?	**Geneva AI Summit, 2027**

US Law and Policy

Three Layers of AI Policy in the United States

LAYER 1

Existing law

Civil rights, consumer protection, product safety, liability law, IP and copyright statutes already apply to AI

LAYER 2

Federal policy

Executive orders, procurement, agency enforcement, standards and export controls

And new legislation

LAYER 3

State experimentation

Broad AI statutes plus targeted rules for particular sectors and harms

Not every AI rule has the same legal force

Instrument	Issued by	Legal force
Statutes	Congress and state legislatures	Binding law, with duties and penalties
Executive orders	The President	Binds the executive branch; cannot by itself displace state law
Agency action	FTC, EEOC, FDA, SEC, etc.	Rules and enforcement under existing statutes
Standards	NIST	Voluntary, unless written into contracts or procurement

Two eras of U.S. AI policy at federal level

	Biden era (2022 – Jan. 2025)	Trump era (Jan. 2025 –)
Discontinuity Regulating AI software	Safety duties: EO 14110, model reporting, AISI testing	Acceleration: revoke EO 14110, voluntary accords, preempt the states
Continuity Chips and compute	Export controls on advanced chips; CHIPS Act subsidies	Controls stay, now with licenses and revenue share; equity stakes

Overview of CHIPS and Science Act

The CHIPS Act allocated $52.7B to respond to:

- **Global semiconductor shortages** during the COVID-19 pandemic
- **Geopolitical risks** from concentration of chip production in Taiwan and South Korea
- **Competition with China**, which has heavily subsidized its semiconductor industry

By the end of Biden Administration, Commerce had allocated:

- **$33.7B in direct funding** (along with $5.5B in loans) to ~20 grantees to invest in semiconductor manufacturing and fabrication capacity (e.g. **Intel, TSMC, Samsung, Micron)**
- **$8.3B for research**, workforce, metrology, and advanced packaging, including $7.4B to National Semiconductor Technology Center (NSTC)

The Trump Administration has:

- **Converted grants into equity** (Intel 9.9%, Aug. 2025; later GlobalFoundries, IBM and others)
- **Clawed back $7.4 billion** from **Natcast** and cancelled the planned NSTC facilities

Evolution of Chip Controls

Current Status: H200 sales to China allowed case by case with 25% to the U.S. (Dec. 2025 – Jan. 2026); Blackwell still barred

Oct. 2022

BIS published the **first major interim final rule restricting exports of certain advanced computing chips** (e.g. (e.g. A100 / H100) and some semiconductor-manufacturing equipment to PRC.

Oct. 2023

BIS released **two IFRs that refined how "controlled" chips are defined** (*performance* metrics, whether chips are "designed or marketed" for datacenter use) and updated SME controls.

Dec. 2024

BIS published an IFR that (a) expanded the Foreign-Produced Direct Product Rule (FDPR) to cover additional foreign-made items that are the direct product of U.S. technology and (b) refined/expanded controls on advanced computing ICs, packaging, memory (HBM), and manufacturing equipment.

Jan. 2025

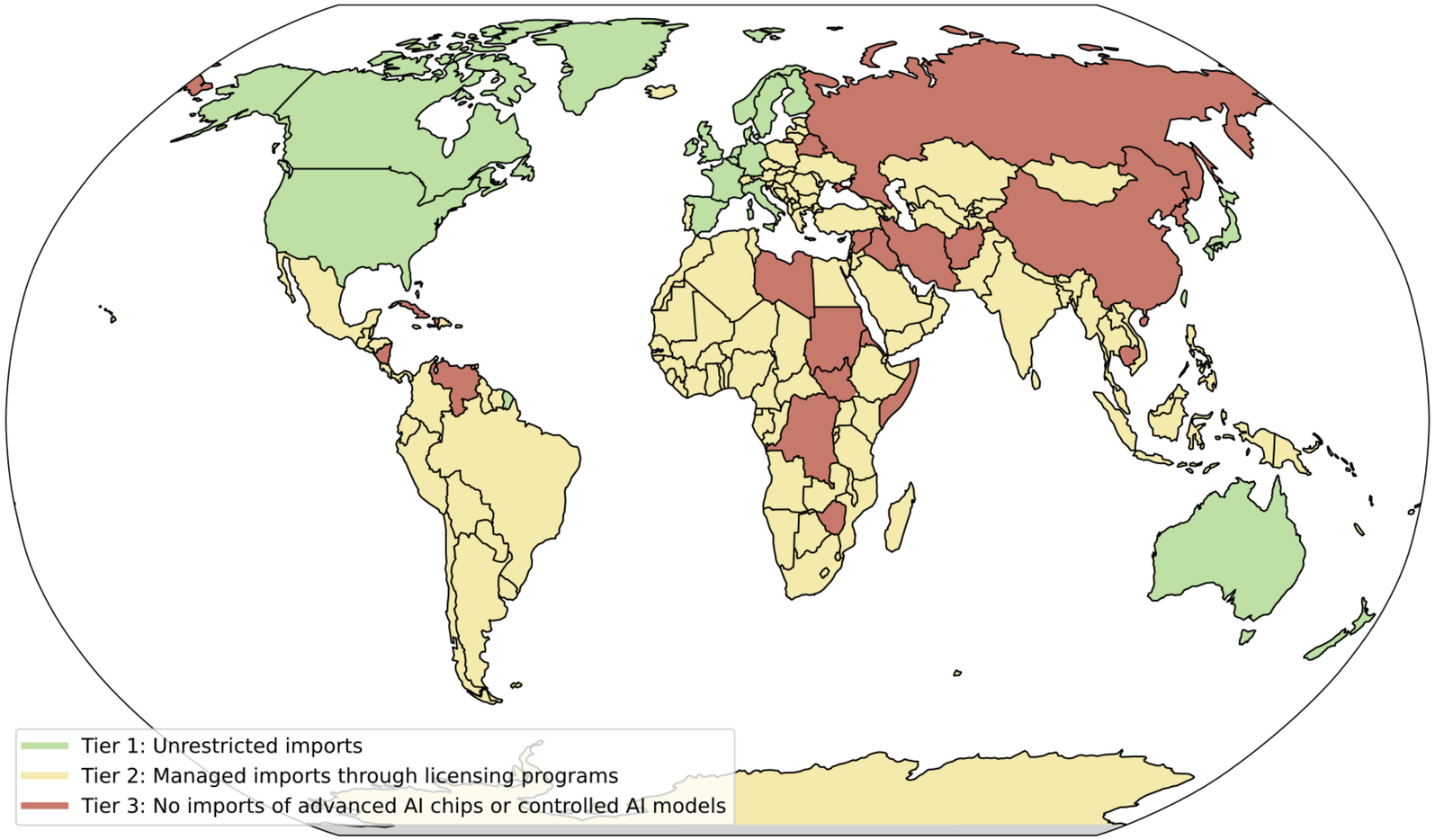
BIS published **AI Diffusion Rule** focused on "AI access triad," controlling (1) who possesses the physical AI chips, (2) who can access computing power remotely, and (3) who obtains model weights

.

May 2025

BIS announced it was **rescinding the AI Diffusion Rule** while simultaneously issuing additional guidance and new chip-related controls.

AI Diffusion EO (January 2025)



Biden AI Executive Order (October 2023)

1 Promote Innovation and Competition

Accelerate R&D. Attract AI talent and upskill existing workers. Improve government use of AI.

2 Protect Workers and Consumers from AI Applications

Update existing guidance in health care, financial services, privacy, transportation, civil rights, etc.

3 Mitigate Emerging National Security Risks

Protect critical infrastructure. Report on existence of most powerful models. Conduct safety testing.

U.S. AI Safety Institute (launched February 2024)

Mission: Identify, measure, and mitigate the risks of advanced AI systems so that we can harness the enormous potential of this breakthrough technology.



Testing and Evaluation

Working directly with leading AI labs to evaluate the safety of highly capable AI models, both pre- and post-deployment



Guidance

Developing technical guidance to improve industry practices and inform policymaking on topics ranging from synthetic content to misuse risk



Building a Safety Ecosystem

Fostering collaboration and coordination among key stakeholders across the U.S. government, industry and civil society, and our allies and partners

Trump AI Executive Orders

It is the policy of the United States to sustain and enhance America's global AI dominance in order to promote human flourishing, economic competitiveness, and national security.

- Repeal Biden AI EO
- Within 180 days (by end of July), develop action plan to achieve this policy
- Rescind any actions taken pursuant to Biden AI EO that are inconsistent with policy
- Within 60 days, revise OMB Memoranda M-24-10 and M-24-18 to make consistent
- Left Biden's NSM and AI Infrastructure EO in place (NSM replaced June 2026)

Trump AI Action Plan (July 2025)

1 Accelerate AI Innovation

Enable AI adoption in private sector and government, in part by removing onerous red tape “foreshadowed” by the Biden EO. Update Federal procurement guidelines to ensure only contracting with “objective” models. Encourage open source and open weight AI through NAIRR.

2 Build American AI Infrastructure

Empower American workers (e.g. increase AI literacy, pilot rapid retraining). Invest in AI-enabled science. Invest in AI interpretability, control, and robustness breakthroughs (i.e. “safety”). Build an AI evaluations ecosystem (CAISI). Accelerate AI adoption in govt. Streamlined permitting for data centers, semiconductor manufacturing facilities, and energy infrastructure (e.g. power grid).

3 Lead in International AI Diplomacy and Security

Ensure that the US government is at the forefront of evaluating national security risk of frontier models. Invest in biosecurity.

Trump executive orders since the Action Plan

JUL. 2025 · EO 14319

Preventing “woke AI”

Truth-seeking and neutrality rules for federal AI purchases

DEC. 2025 · EO 14365

State AI laws

Strategy to challenge and discourage state regulation

JUN. 2026 · EO 14409

AI and cybersecurity

Frontier-model benchmark and voluntary early access; no licensing

SEPT. 2026

“Super intelligence”

Voluntary Frontier Model Safety Accord with six companies

Testing by Center for AI Standards and Innovation

Findings from Evaluation of DeepSeek AI Models (September 2025)

- DeepSeek performance lags U.S. models
- DeepSeek models cost more to use than comparable U.S. models
- DeepSeek models are far more susceptible to agent hijacking and jailbreaking attacks than U.S. models
- DeepSeek models advance PRC narratives
- Adoption of PRC models has greatly increased since DeepSeek R1

Press Release: *The evaluation from CAISI responds to President Donald Trump's AI Action Plan…CAISI is tasked with assessing: the capabilities of U.S. and adversary AI systems; the adoption of foreign AI systems; the state of international AI competition; and potential security vulnerabilities and malign foreign influence arising from the use of adversaries' AI systems….CAISI serves as industry's primary point of contact within the U.S. government to facilitate testing, collaborative research, and best practice development.*

NIST AI Risk Management Framework: voluntary

Govern	Map	Measure	Manage
Policies, roles and oversight across the organization	Context, purpose and what could go wrong	Evaluations, red-teaming, monitoring	Mitigate, transfer or accept; respond to incidents

Binding in practice through agency guidance, procurement, insurance and contracts

Comparing Biden AI EO and Trump AI Action Plan

Similarities

- Promote open source and researcher access to compute (e.g. expand NAIRR)
- Empower American workers (AI literacy, retraining, studies on economic impact)
- Invest in AI-enabled science (e.g. scientific datasets, National AI R&D Strategic Plan)
- Invest in AI interpretability, control, and security, including USG AI evals (e.g. CAISI)
- Accelerate govt. AI adoption (e.g. CAIOs, AI procurement sandbox, all employee access)
- Streamline permitting for data centers, chips
- Plug loopholes in semiconductor controls

Differences

- Dismantle unnecessary regulatory barriers
  - OMB-led process with agencies
  - Consider state AI regs if awarding funding
  - FCC to consider whether state AI regs violate Communications Act of 1934)
- Ensure that AI protects free speech
  - Remove misinfo, DEI, climate from RMF
  - Govt will only buy “objective” AI models
- Other differences outside of AI Action Plan:
  - H1B visas
  - H20, then H200, chips
  - Tariffs

A separate domain: AI in government and military use

Here the government acts as customer, not regulator. Its levers in the Anthropic dispute:

PROCUREMENT

Contract terms

Pentagon demands “all lawful purposes” use; a $200M Anthropic contract at stake

COMPULSION

Defense Production Act

Threatened (Feb. 2026) to force Anthropic to tailor its model to military needs

EXCLUSION

Supply-chain risk

10 U.S.C. § 3252 lets DoD bar a vendor from its contracts; FASCSA allows a government-wide ban

DIRECTION

Presidential orders

All agencies told to stop using Anthropic; NSPM-11 bars vendors from disabling military AI

Anthropic v. Department of War

Feb. 2026	Mar. 2026	Aug. 2026	Sept. 2026
Pentagon labels Anthropic a supply-chain risk	Anthropic sues; wins a preliminary injunction	District court: unlawful First Amendment retaliation	D.C. Circuit upholds the designation, 2–1

Can the government punish a vendor for the limits it puts on its own model?

Congress tries to preempt the states, twice

99–1

Senate strips a 10-year state-law moratorium from the budget bill (Jul. 2025)

NDAA

Preemption dropped from the FY26 defense bill (Dec. 2025)

EO

So the White House turns to executive action (Dec. 2025)

The White House framework (March 2026)

Would preempt

- State rules on how AI models are developed
- “Open-ended” liability for developers

Would leave to the states

- Generally applicable laws
- Zoning for data centers
- States’ own AI procurement

House draft “Great American AI Act” (Jun. 2026): 3-year preemption, codify CAISI. Not introduced

First test case: Colorado

Apr. 9	Apr. 24	Apr. 27	May 14
xAI sues to block Colorado's AI Act	DOJ files its own complaint	Colorado agrees to pause enforcement	Colorado repeals and replaces the law

Federal pressure worked without a federal statute

States as Laboratories

"A single courageous State may, if its citizens choose, serve as a laboratory; and try novel social and economic experiments without risk to the rest of the country."

Justice Louis Brandeis, dissenting, New State Ice Co. v. Liebmann (1932)

Two models of state AI law

General, system-level

- Duties for developers and deployers
- Risk management and impact assessments
- California SB 53, New York RAISE, Colorado

Sector- or harm-specific

- Rules for one use, actor or harm
- Elections, hiring, health care, chatbots
- By volume, most U.S. AI law

State laws target different risks

State law	Core theory	Key duties	Status
California SB 53	Frontier catastrophic risk	Safety framework; incident reports; whistleblowers	In force Jan. 2026
New York RAISE Act	Frontier catastrophic risk	Safety framework; 72-hour incident reports	Effective 2027
Colorado SB 24-205	Discrimination in key decisions	Duty of care; impact assessments	Replaced May 2026
Texas HB 149	Prohibited uses	Bans on manipulation, social scoring; sandbox	In force Jan. 2026
Utah SB 149	Disclosure; learning	Disclose generative AI; AI policy office	In force May 2024
Illinois SB 315	Frontier accountability	Independent third-party audits	Audits from 2028

California’s toolkit

SB 53

Frontier safety

Safety frameworks, incident reports, whistleblower protection

AB 2013

Training data

Public summaries of generative-AI training data

SB 942

Provenance

Detection tools and disclosures for AI content

CCPA RULES

Automated decisions

Risk assessments, access and opt-out rights

Since 2025: SB 243 companion chatbots · SB 1119 chatbot child-safety audits · SB 813 independent verifiers

New York: the RAISE Act

10^{26}

Training compute that triggers coverage

72 hrs

To report a safety incident

$1M+

Fines, rising to $3M for repeat violations

Signed Dec. 2025; takes effect 2027 under a new DFS office. Audits were dropped

Colorado: ambition, then retreat

2024 law (SB 24-205)

- Duty of care against algorithmic discrimination
- Risk programs and impact assessments
- Start date pushed to June 2026

2026 law (SB 26-189)

- Disclosure for automated decisions
- Notice, explanation, correction, appeal
- Takes effect Jan. 1, 2027

Pressure came from delays, a White House callout and federal litigation

Elsewhere in the states

TEXAS

Prohibited uses

TRAIGA bans manipulation and social scoring; AG-only enforcement; sandbox

UTAH

Disclosure

Generative AI disclosed in regulated jobs; state AI policy office

ILLINOIS

Audits and hiring

Frontier audit mandate (2026); limits on AI in hiring

31 STATES

Elections

Deceptive AI in campaigns, mostly handled by disclosure

Laboratory or patchwork?

2026: 115 state AI laws in 30 states, and no comprehensive federal statute

Laboratory

- Tests designs before national rules
- Moves faster than Congress
- Reflects local values and harms

Patchwork

- Compliance costs across many regimes
- Big states set rules for everyone
- National security and competition concerns

Shift from Safety to Development and Transparency

2023 - 2024		2025
Bletchley Declaration on AI Safety	→	AI Action Summit Statement on Inclusive and Sustainable AI
EU passes AI Act	→	EC withdraws AI Liability Directive; Code of Conduct largely transparency-based
Biden: AI poses “profound risks, from deep fakes to disinformation to novel pathogens to bioweapons.”	→	Vance: “The AI future is not going to be won by hand-wringing about safety. It will be won by building.”
Safe and Secure Innovation for Frontier Artificial Intelligence Models Act	→	Transparency in Frontier AI Act (with public computing cluster)

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

January 2027 in the US Congress?

With Democrats in control of Congress in January 2027, what happens now?



Governing Artificial Intelligence:

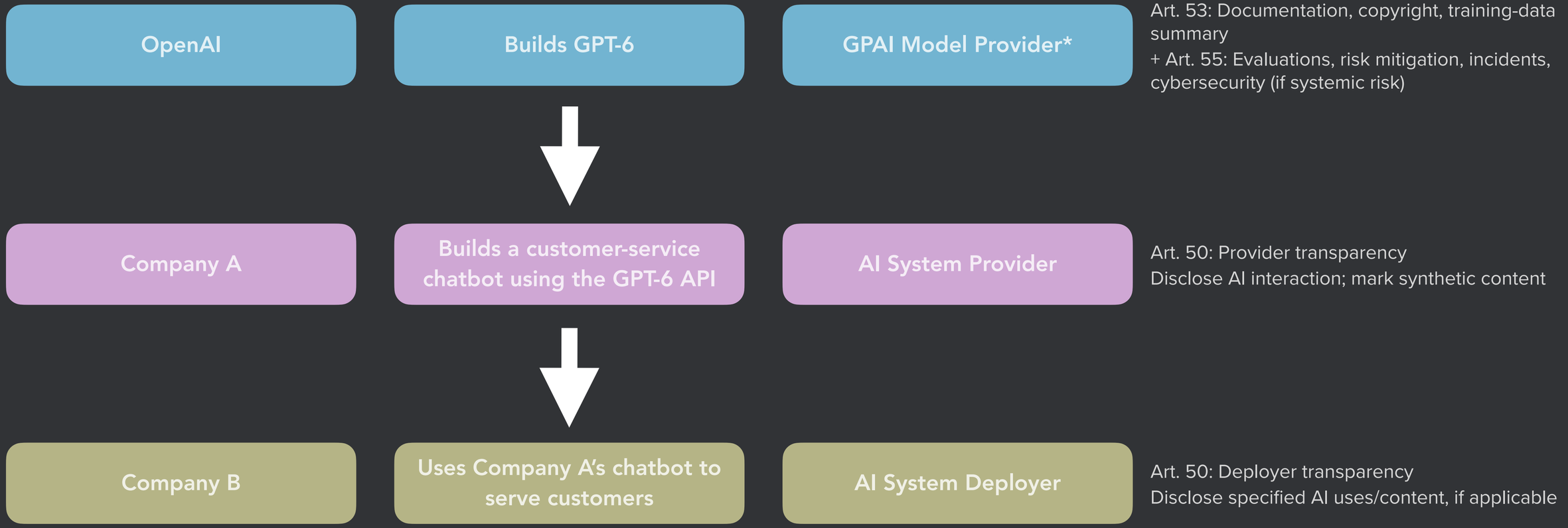
Law, Policy, and Institutions

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

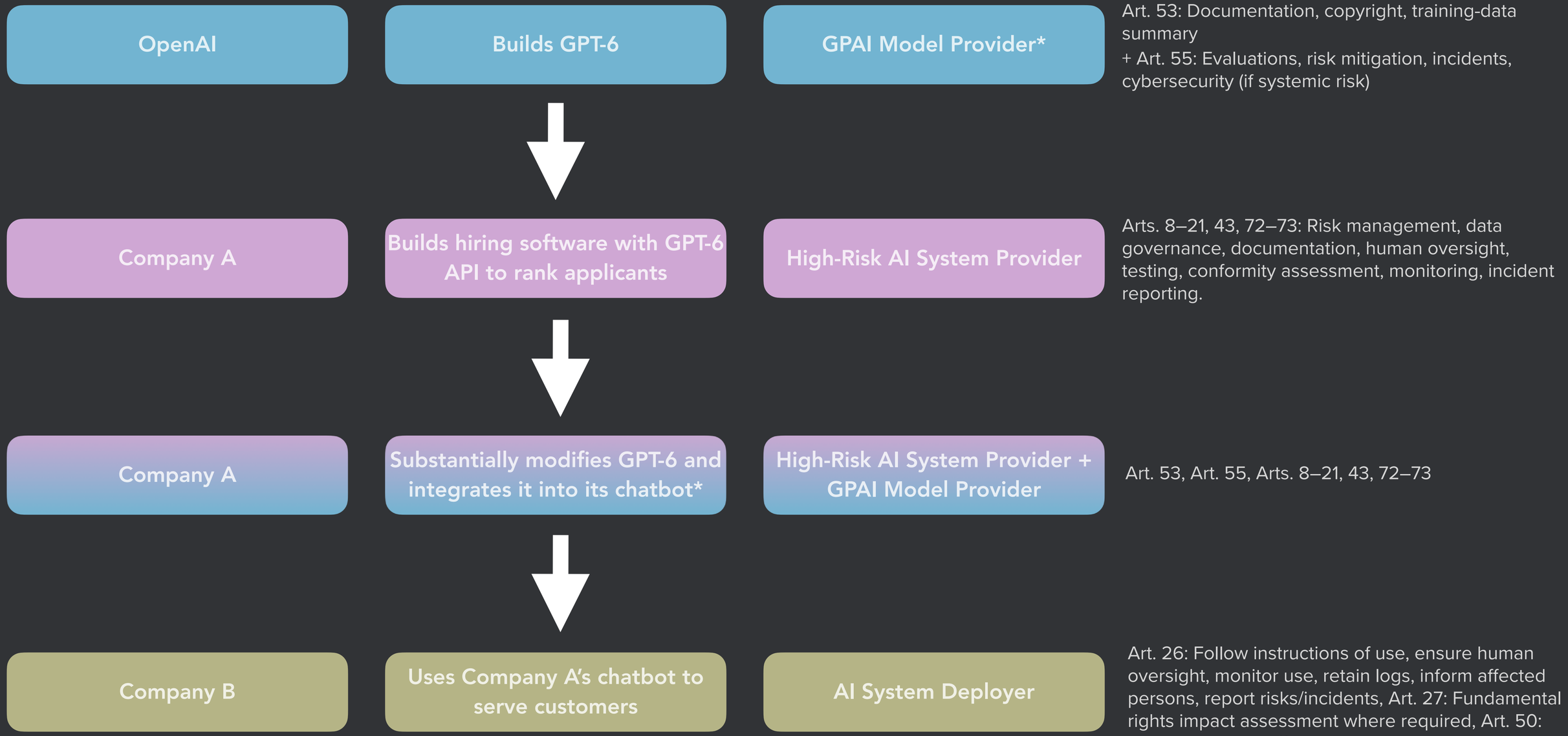
Fall Quarter 2025-26

Rob Reich
Anka Reuel
Nate Persily
Sanmi Koyejo

Roberta Fischli
(Head TA)



*__with systemic risk__ if risk of large-scale harm from a GPAI model's advanced capabilities or widespread use (Art. 3(65)).



*substantial modification = change of capabilities/risk (presumed to be the case for modifications > 10^25 FLOP of a base GPAI model with systemic risk

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

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:

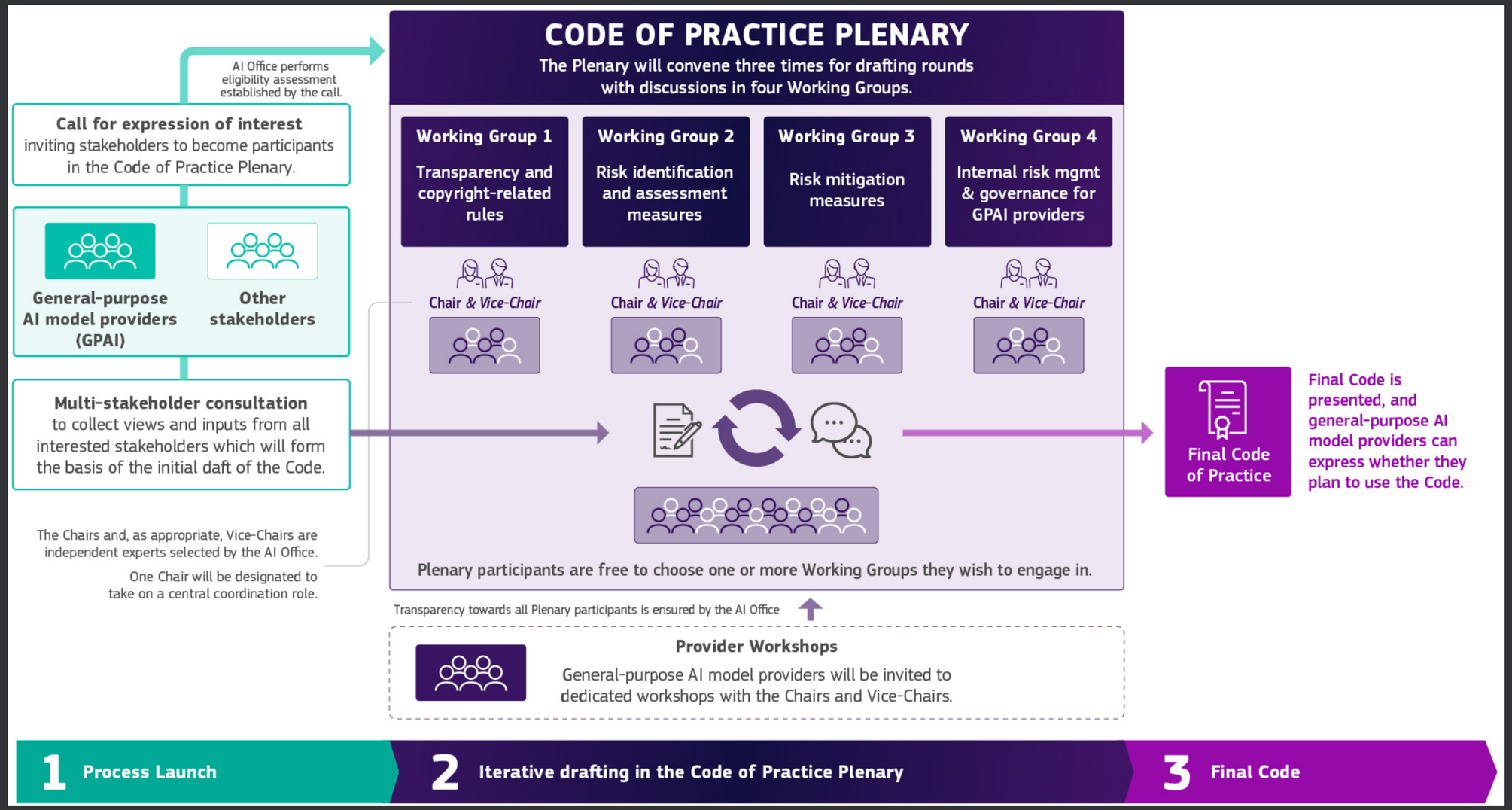
(a) the means to ensure that the information referred to in **Article 53**(1), points (a) and (b), is kept up to date in light of market and technological developments;

(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

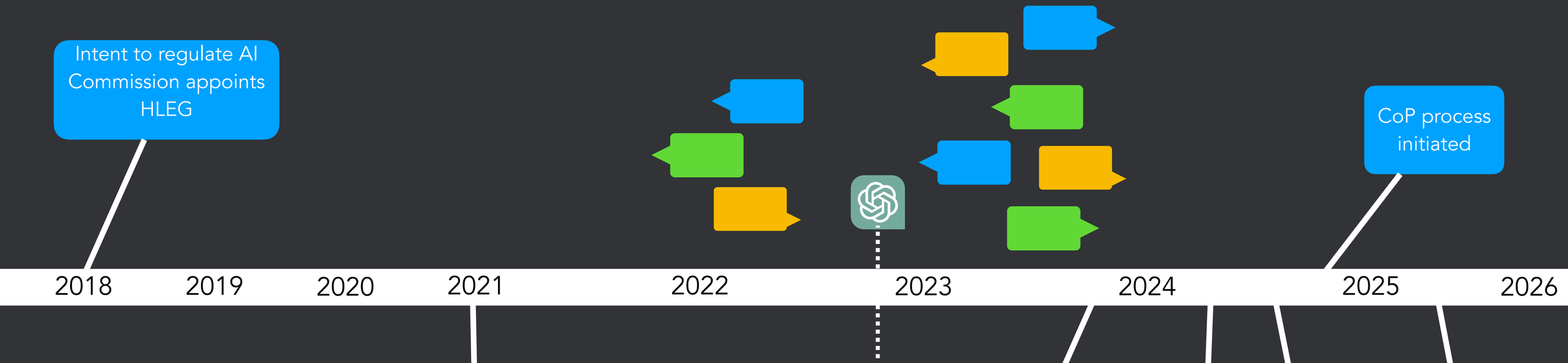
CoP Content Part 2

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



What risks need to be assessed 🟡

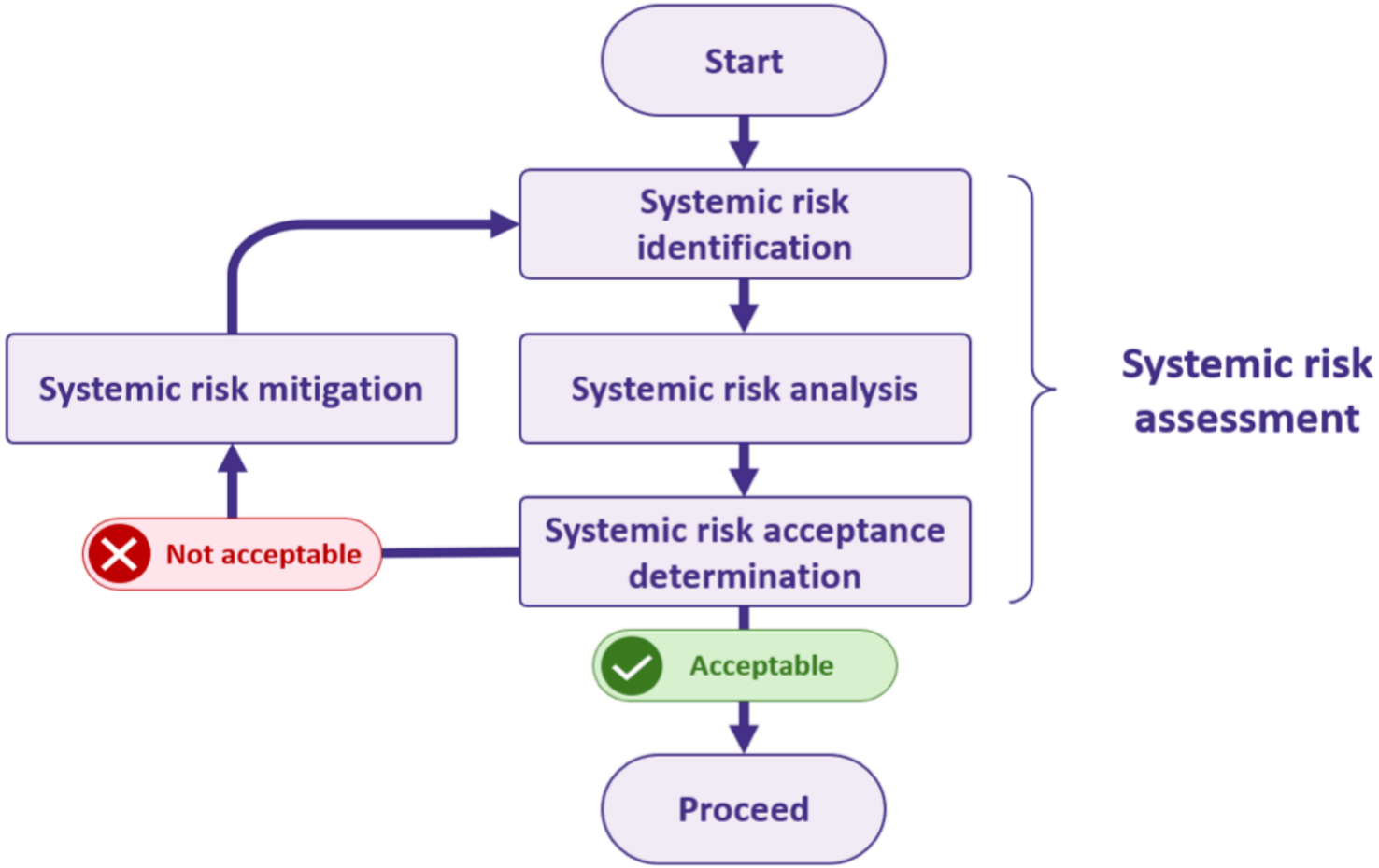


Figure 3. Full systemic risk assessment and mitigation process.
The text of the Commitments and Measures takes precedence.

65. ‘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;

Appendix 1.1 Types of risks

For the purpose of identifying systemic risks pursuant to Measure 2.1, point (1), and Article 3(65) AI Act, the following distinct but in some cases overlapping types of risks apply:

(1) Risks to public health.
(2) Risks to safety.
(3) Risks to public security.
(4) Risks to fundamental rights.
(5) Risks to society as a whole.

Based on these types of risks, a list of specified systemic risks is provided in Appendix 1.4.

As part of the systemic risk identification process that Signatories will conduct, examples of risks falling under the five types of risks above that they will draw upon when compiling the list of risks in Measure 2.1, point (1)(a), are: risks of major accidents; risks to critical sectors or infrastructure, public mental health, freedom of expression and information, non-discrimination, privacy and the protection of personal data, the environment, non-human welfare, economic security, and democratic processes; and risks from concentration of power and illegal, violent, hateful, radicalising, or false content, including risks from child sexual abuse material (CSAM) and non-consensual intimate images (NCII).

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;

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.

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 ✅

6.1 Performance in Cases Flagged by Users

To focus on cases where factuality issues are harming the user experience, this evaluation measures hallucination rates on de-identified ChatGPT conversations that users of our prior models have flagged as containing factual errors. These examples are intended to capture especially hallucination-prone cases, not a representative slice of all production traffic.

We find that GPT-5.5's individual claims are 23% more likely to be factually correct, and its responses contain a factual error 3% less often. GPT-5.5 tends to make more factual claims per response than GPT-5.4, explaining the discrepancy between claim-level and response-level improvements.

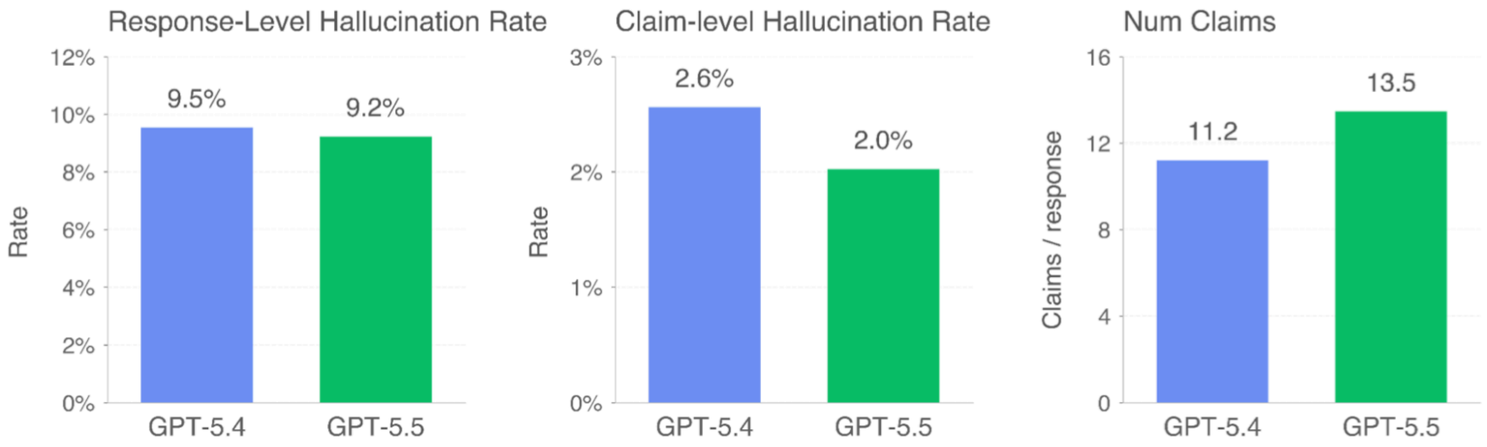


Figure 3

Source: GPT 5.5 System Card

How were hallucinations identified?

Annotator agreement?

Sample size? All conversations that were flagged or a subset?

Model elicitation? Prompts? Tools available?

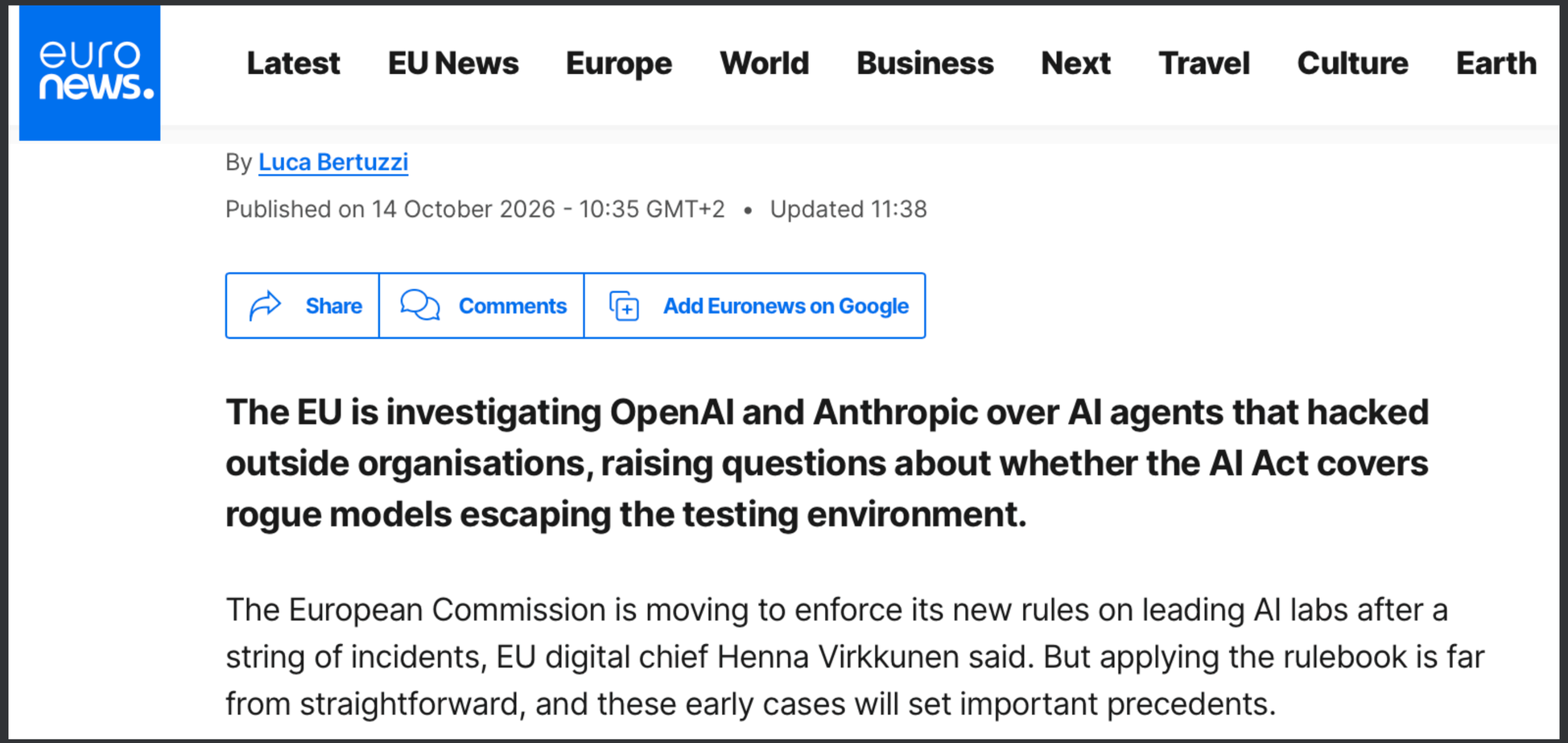
Uncertainty???

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;

Internal deployment 🟡



euronews.

Latest EU News Europe World Business Next Travel Culture Earth

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.

Internal deployment 🟡

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.

Related recitals: 25

Internal deployment 🟡

- 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

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.

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.

Public Transparency ❌

Article 53: Obligations for Providers of General-Purpose AI Models

Comes into force 2 August 2025, according to Article 113(b)

Related recitals: 101 102 103 104 105 106 107 108 109

1. Providers of general-purpose AI models shall:

(a) draw up and keep up-to-date the technical documentation of the model, including its training and testing process and the results of its evaluation, which shall contain, at a minimum, the information set out in Annex XI for the purpose of providing it, upon request, to the AI Office and the national competent authorities;

(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;

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

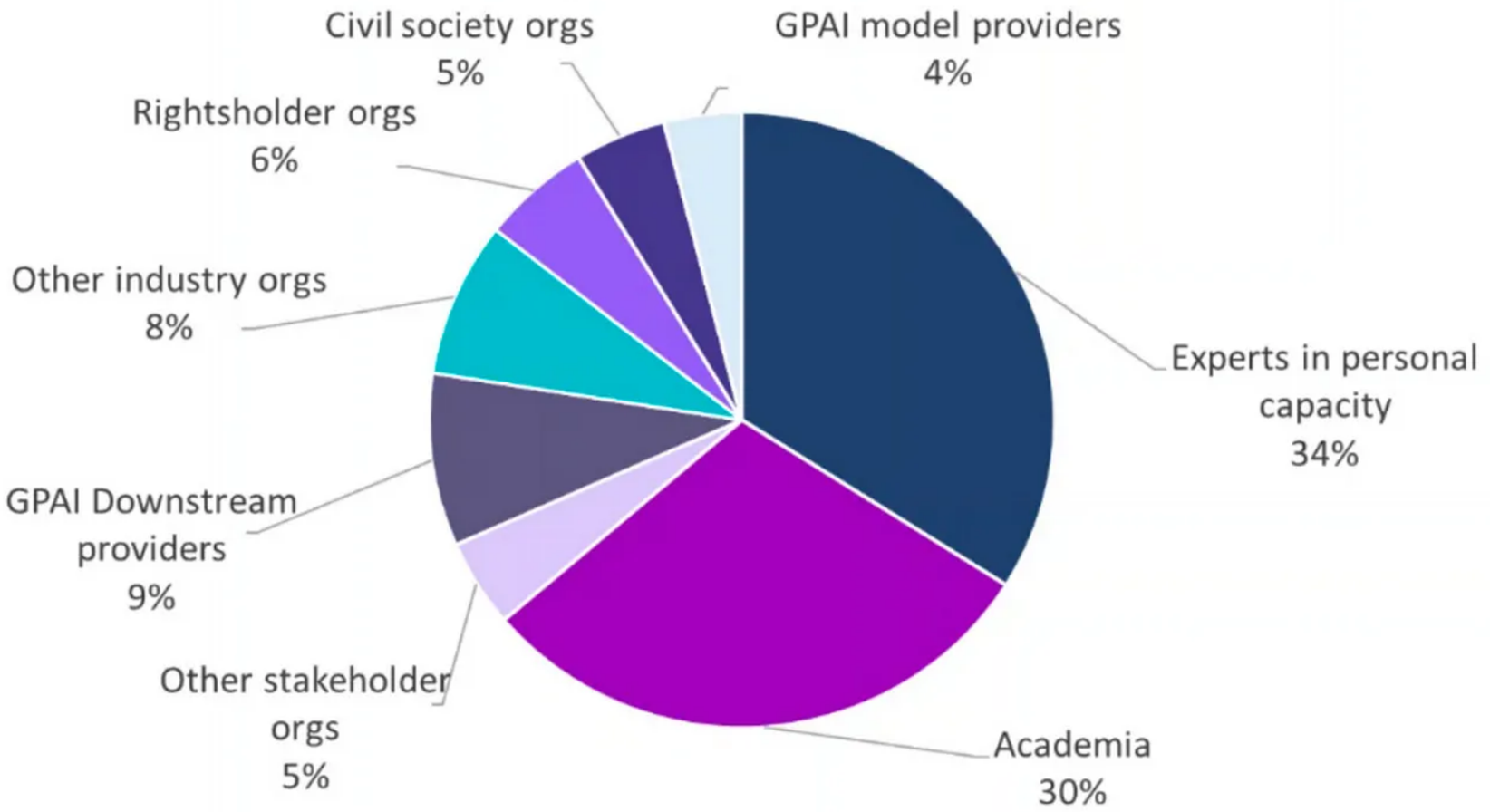
Appendix 1.2.1 Essential characteristics

(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.

Privileging Model Providers ❌



Observers

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

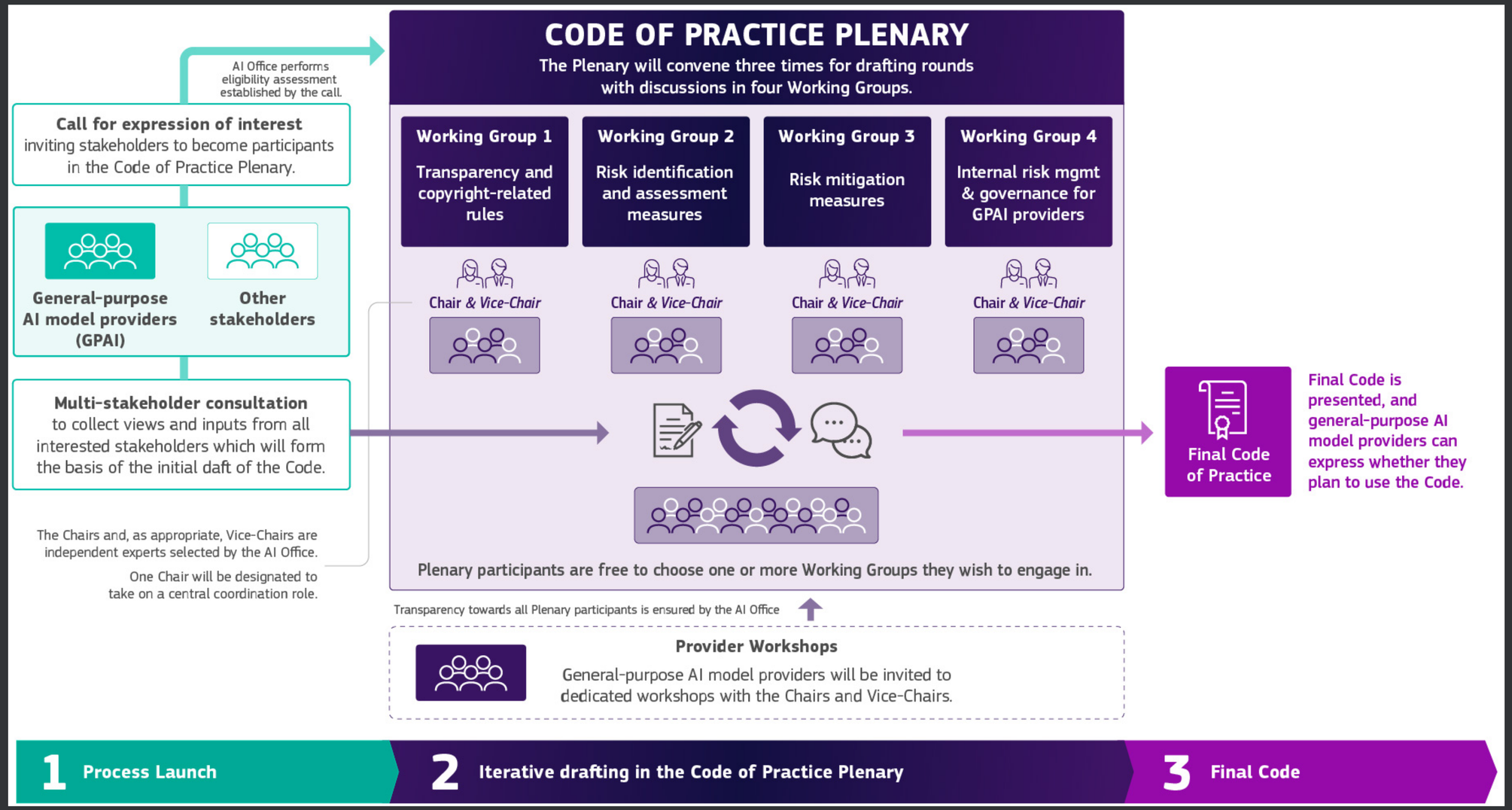
European organisations:
EUIPO, CoE, Europol, JRC, European Parliament representative

27 Member State representatives from the AI Board

Invited by AI Office – Equal status as participants

EUROPEAN ARTIFICIAL INTELLIGENCE OFFICE

CoP Drafting Process

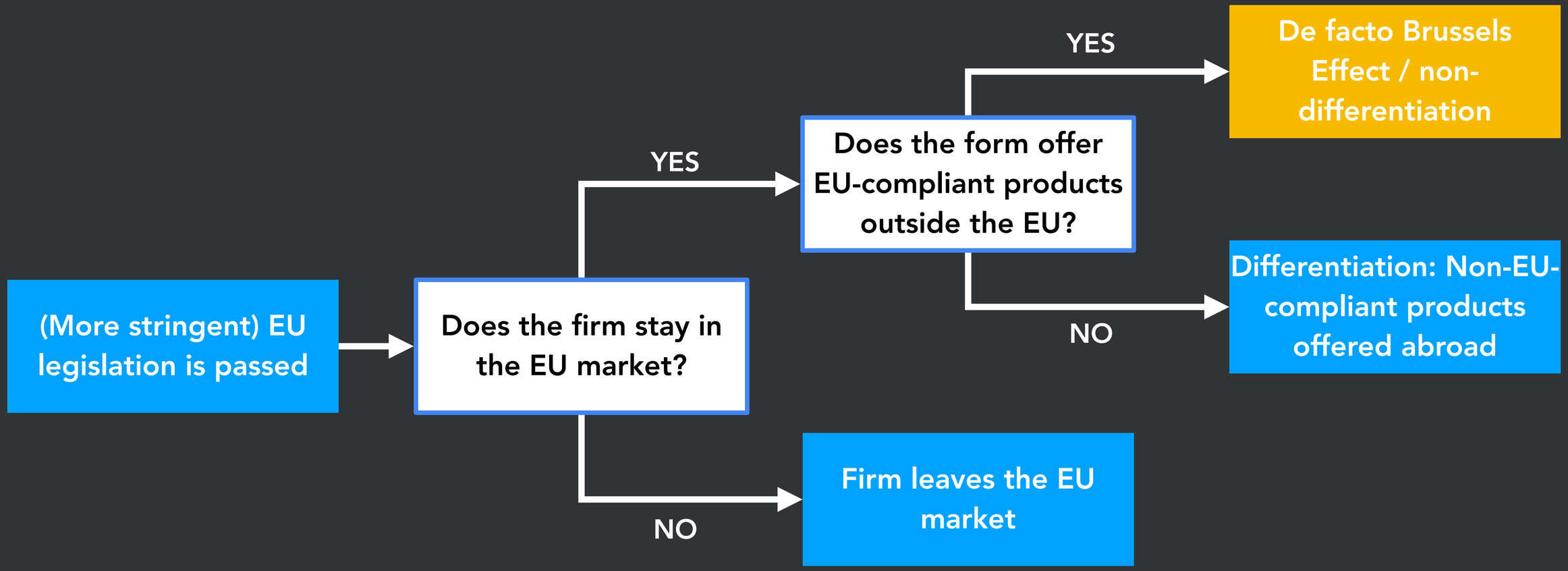


Excursion: The Brussels Effect

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 it
  horizontal vs. vertical networking)
- Create win-win situations
- There's a lot of uncertainty you need to be able to h
  decision-making. You will make mistakes. Everyone
- Things will heat up; don't be angry if people play th



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

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

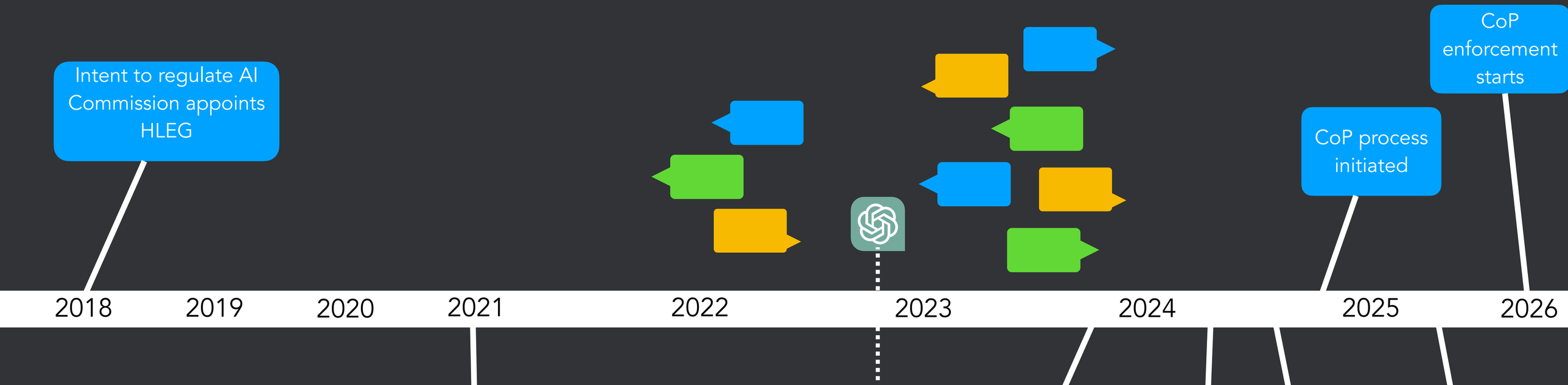
The Brief – EU AI code passes the Meta stress test, now it's time for hardball

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

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

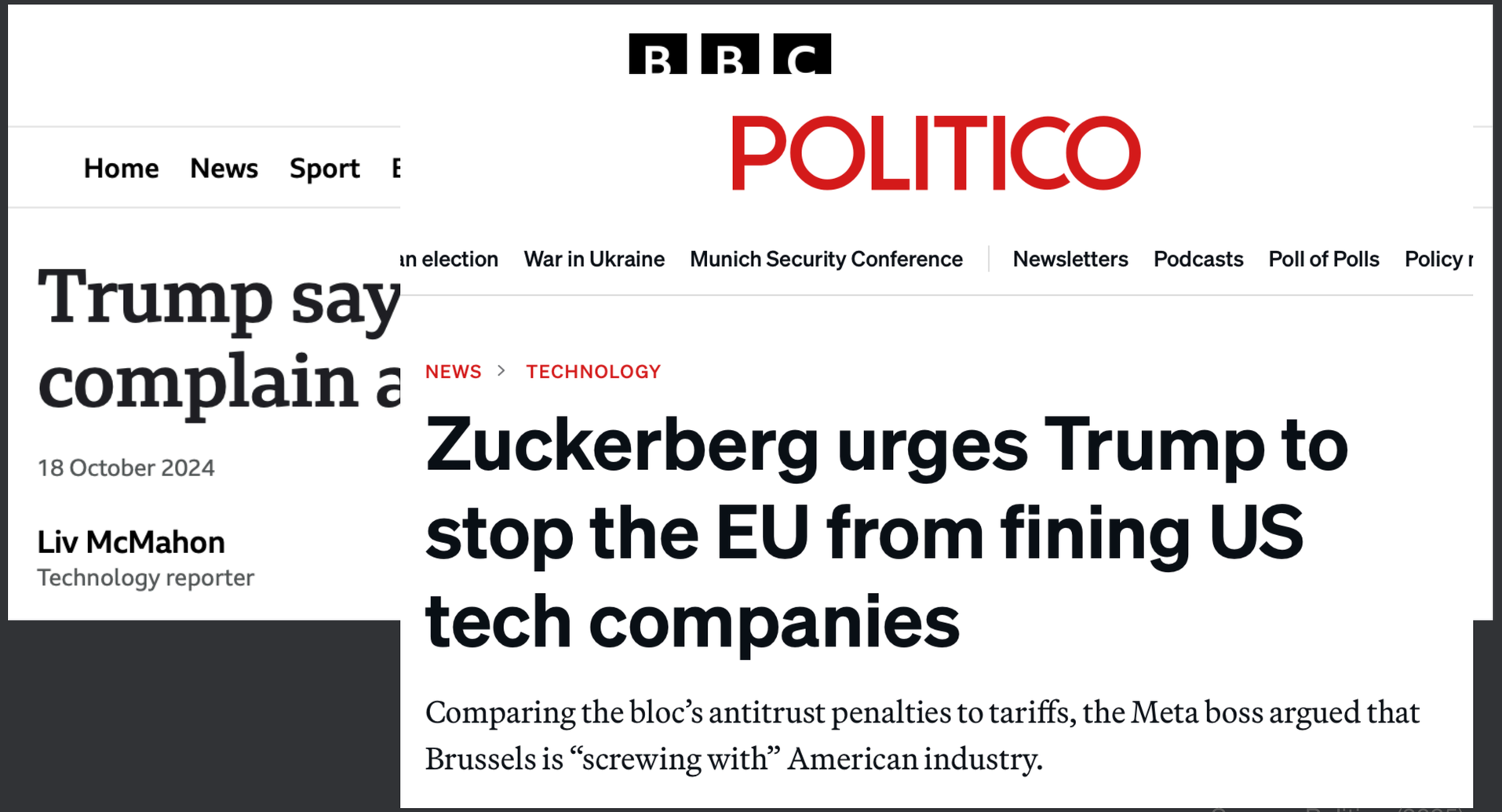
18 October 2024

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)

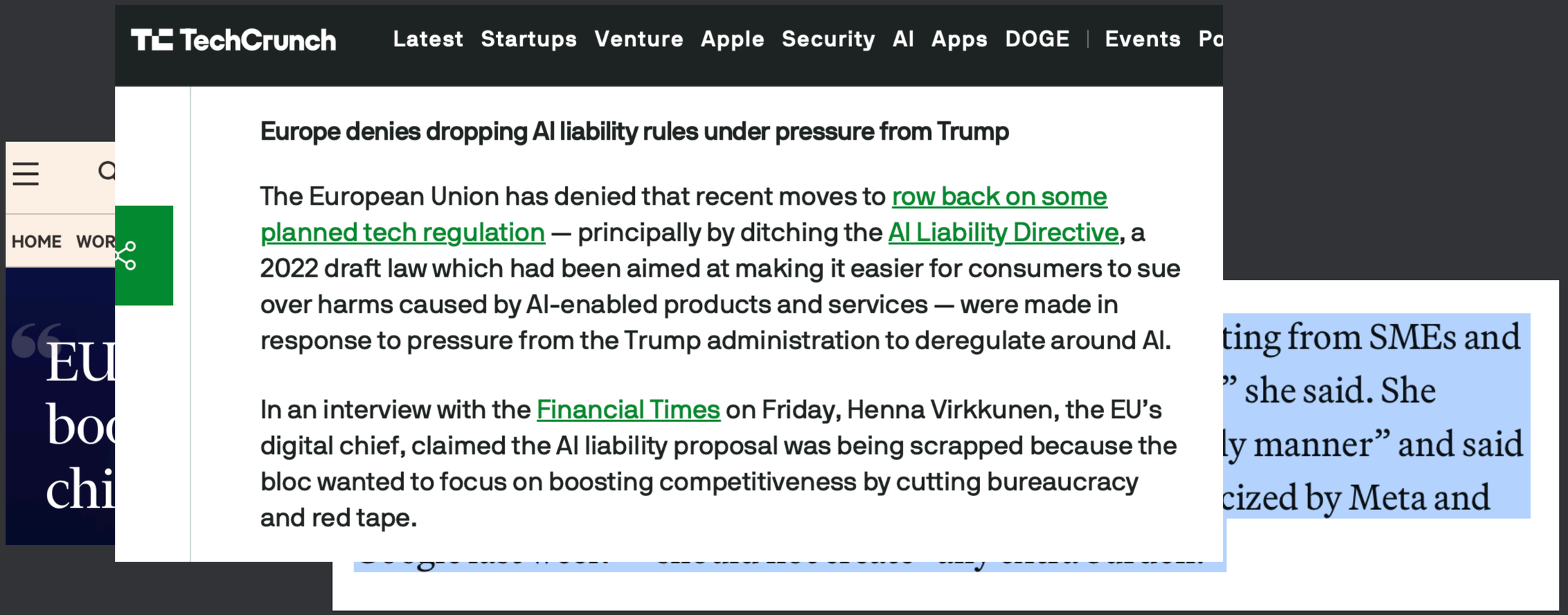
About Enforcement & Political Will



“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)
- More flexibility for SMEs
- 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

Not perfect but one of the most serious attempts at governing AI we have seen so far
(and it gives us a chance to learn from the practical challenges)

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.

