

**Road to Amherst**

# Update on AI regulation

IEEE Emerging Technology  
Reliability Roundtable  
May 21-22, 2024

Lynette Webb  
lynette@roadtoamherst.com

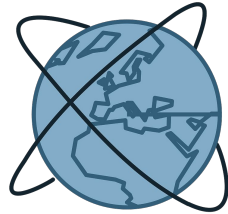


*Image generated by DALL-E*

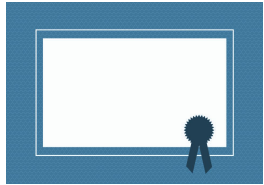
# Key players in AI policy



**Governments and regulators** set the legal frameworks that govern the development and deployment of AI technologies



**Intergovernmental bodies** such as the OECD, UNESCO and G7 are active forums for AI policy discussion



**Standards bodies** like ISO, IEEE (global) and NIST, ESOs (regional) offer guidance and performance yardsticks



**Academic and research institutions** are part of technical and societal debates on AI



**Lobbyists and think tanks** promote industry or civil society viewpoints

# Key areas of concern

## Operational practices

**Risk of unfair bias/discrimination** if AI systems are used for profiling or decision-making - “How can we prevent AI systematically discriminating against certain people?”

**Use of data** for model training without permission/control/compensation”

- Copyrighted materials - “Shouldn’t AI developers pay content owners for the data they train on?”
- Personal/private data - “Is AI violating people’s privacy?”

**Environmental impact** - “Is AI making climate change worse?”

**Lack of transparency / human oversight** - “How can we be confident AI is safe/accurate if it’s a black box that no one is checking? What are the remedies if an AI system’s output is wrong?”

- Explainability of AI system output
- Accountability and redress mechanisms
- Awareness when AI systems are being used

## Wider impact on society

**Disruption to jobs and employment shifts** - “Will AI put millions of people out of work?”

**Misuse of AI by bad actors** - “Is AI giving criminals better weapons?”

- Misinformation
- Surveillance
- Hacking/fraud
- New weapons (e.g., chemical, killer robots)

**Problem of AI alignment** (worsened by rapid acceleration in capabilities) - “Is what an AI wants aligned with what humans want?”

**Shifting balance of power** - “Is AI undermining democracy? Will global stability be disrupted?”

- Big Tech monopolisation
- Geopolitical incl military (e.g., US vs China vs Russia)

← **Unethical manipulation by using AI for microtargeting** →  
- “Is AI undermining people’s free will?”



# 1. Recent milestones in AI regulation

2. Deeper dives into key legislation:

- **US Executive order** on safe, secure and trustworthy development and use of AI
- **European AI Act**



## 1. Recent milestones in AI regulation

2. Deeper dives into key legislation:

- **US Executive order** on safe, secure and trustworthy development and use of AI
- **European AI Act**



# Recent milestones in AI regulation

US and Europe  
Oct 2023 – Apr 2024

*Key changes since last year's summit:*

- Increased focus on 'foundation models' and broader AI safety issues beyond fairness/data – driven initially by UK
- The US is no longer on the sidelines, and is leading by example with AI regulation for Federal agencies



Oct

30 October 2023: G7 publishes ‘Hiroshima Process’  
Guiding Principles for developing Advanced AI Systems  
and a Code of Conduct for developers

*30 October 2023:* **President Biden issues**  
**Executive Order on Safe, Secure, and**  
**Trustworthy Development and Use of AI**

Oct

*30 October 2023:* G7 publishes ‘Hiroshima Process’  
**Guiding Principles for developing Advanced AI Systems**  
and a **Code of Conduct** for developers



# Summary – US Executive Order on AI

150+ actions spanning 50+ federal entities

Ambitious deadlines – 69 actions were due within 180 days (by end April 2024)

Introduced reporting requirements for foundation models and large-scale compute capacity

1. Purpose
2. Policy and Principles
3. Definitions
4. Ensuring the Safety and Security of AI Technology
5. Promoting innovation and competition
6. Supporting workers
7. Advancing Equity and Civil Rights
8. Protecting Consumers, Patients, Passengers and Students
9. Protecting Privacy
10. Advancing Federal Government Use of AI
11. Strengthening American Leadership Abroad
12. Implementation
13. General provisions

30 October 2023:  
President Biden  
issues Executive  
Order on Safe,  
Secure, and  
Trustworthy  
Development and  
Use of AI

*1-2 November 2023: UK stages global  
AI safety summit, culminating in the  
Bletchley Declaration*

Nov

30 October 2023:  
G7 publishes  
'Hiroshima  
Process' Guiding  
Principles for  
developing  
Advanced AI  
Systems

# The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023

... To inform action at the national and international levels, our agenda for addressing frontier AI risk will focus on:

- **identifying AI safety risks of shared concern**, building a shared scientific and evidence-based understanding of these risks...
- **building respective risk-based policies across our countries to ensure safety** in light of such risks, collaborating as appropriate while recognising our approaches may differ...

... We resolve to support an internationally inclusive network of scientific research on frontier AI safety...

Australia	Saudi Arabia
Brazil	Netherlands
Canada	Nigeria
Chile	The Philippines
China	Republic of Korea
European Union	Rwanda
France	Singapore
Germany	Spain
India	Switzerland
Indonesia	Türkiye
Ireland	Ukraine
Israel	United Arab Emirates
Italy	United Kingdom
Japan	United States
Kenya	

30 October 2023:  
President Biden  
issues Executive  
Order on Safe,  
Secure, and  
Trustworthy  
Development and  
Use of AI

*1-2 November 2023: UK stages global  
AI safety summit, culminating in the  
Bletchley Declaration*

Nov

30 October 2023:  
G7 publishes  
'Hiroshima  
Process' Guiding  
Principles for  
developing  
Advanced AI  
Systems

*1 November 2023: UK Frontier AI taskforce  
reconstituted as UK AI safety institute*

*1 November 2023: NIST announces it will  
establish the US AI Safety Institute*

30 October 2023:  
President Biden  
issues Executive  
Order on Safe,  
Secure, and  
Trustworthy  
Development and  
Use of AI

1-2 November  
2023: UK stages  
global AI safety  
summit,  
culminating in  
the Bletchley  
Declaration

Dec

30 October 2023:  
G7 publishes  
'Hiroshima  
Process' Guiding  
Principles for  
developing  
Advanced AI  
Systems

1 November 2023:  
UK Frontier AI  
taskforce  
reconstituted as  
UK AI safety  
institute

1 November 2023:  
NIST announces it  
will establish the  
US AI Safety  
Institute

*18 December 2023:*  
**ISO/IEC 42001**  
**AI management system  
standard is published**

- ISO/IEC 42001 is the **first international standard for implementing AI responsibly that companies can 'certify' against.**
- Topics included: risk management, AI system impact assessment, system lifecycle management and third-party suppliers.

30 October 2023: President Biden issues [Executive Order on Safe, Secure, and Trustworthy Development and Use of AI](#)

1-2 November 2023: UK stages global AI safety summit, culminating in the [Bletchley Declaration](#)

In January, 211 AI-related bills were introduced by [US State legislators](#); 101 relating to deep fakes. As of February, 407 AI-related bills were under consideration

Jan

30 October 2023: G7 publishes 'Hiroshima Process' [Guiding Principles for developing Advanced AI Systems](#)

1 November 2023: UK Frontier AI taskforce reconstituted as [UK AI safety institute](#)

18 December 2023: [ISO/IEC 42001](#) AI management system standard is published

1 November 2023: NIST announces it will establish the [US AI Safety Institute](#)

30 October 2023: President Biden issues [Executive Order on Safe, Secure, and Trustworthy Development and Use of AI](#)

1-2 November 2023: UK stages global AI safety summit, culminating in the [Bletchley Declaration](#)

In January, 211 AI-related bills were introduced by [US State legislators](#); 101 relating to deepfakes. As of February, 407 AI-related bills were under consideration

Feb

**6 February 2024: UK published its [response to the AI regulation consultation](#). Plan is to take a context-specific approach led by sector regulators, supported by central oversight/coordination**

30 October 2023: G7 publishes 'Hiroshima Process' [Guiding Principles for developing Advanced AI Systems](#)

1 November 2023: UK Frontier AI taskforce reconstituted as [UK AI safety institute](#)

18 December 2023: [ISO/IEC 42001](#) AI management system standard is published

1 November 2023: NIST announces it will establish the [US AI Safety Institute](#)

*13 March 2024:*  
**European Parliament  
approves final text of  
AI Act**



Mar



# Summary – EU AI Act

- **Prohibitions** on certain uses of AI
- **Mandatory requirements for “high risk AI systems”, “general purpose AI models”** and some other narrow applications
  - Key emphasis is on risk/impact assessment and transparency
  - Differing obligations for providers vs deployers
  - Exemptions for R&D, most open source, some ‘grandfathered’ products, law enforcement/defence
  - Some flexibility for products already subject to regulation
- **Enforcement via large fines and new oversight bodies** including Office of AI and market surveillance authorities.
- Staggered **implementation from end 2024-2026**

14 March 2024: **Council of Europe finalises draft of 'world's first treaty on AI'** focusing on human rights; but it's left up to individual countries whether to include defense and private sector activity within scope

13 March 2024:  
**European Parliament approves final text of AI Act**

Mar

14 March 2024: **Council of Europe finalises draft of 'world's first treaty on AI'** focusing on human rights; but it's left up to individual countries whether to include defense and private sector activity within scope

13 March 2024:  
**European Parliament approves final text of AI Act**

Mar


21 March 2024: **UN General Assembly passed a US-led resolution** promoting the development of safe and trustworthy AI to help meet the UN's sustainable development goals.



# 1. Recent milestones in AI regulation

2. Deeper dives into key legislation:

- **US Executive order** on safe, secure and trustworthy development and use of AI
- **European AI Act**



# **US Executive order** on safe, secure and trustworthy development and use of AI

## *Key points*

- The US has seized the lead in setting standards for AI responsibility – in particular for ‘foundation models’.
- The bulk of the Executive Order is focused on Federal Agencies rather than private companies – but will have wider influence

# Summary – US Executive Order on AI

150+ actions spanning 50+ federal entities

Ambitious deadlines – 69 actions were due within 180 days (by end April 2024)

Introduced reporting requirements for foundation models and large-scale compute capacity

1. Purpose
2. Policy and Principles
3. Definitions
4. Ensuring the Safety and Security of AI Technology
5. Promoting innovation and competition
6. Supporting workers
7. Advancing Equity and Civil Rights
8. Protecting Consumers, Patients, Passengers and Students
9. Protecting Privacy
10. Advancing Federal Government Use of AI
11. Strengthening American Leadership Abroad
12. Implementation
13. General provisions

## Section 1: Purpose ([link](#))

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

## Section 2: Policy and Principles ([link](#))

- A. AI must be safe and secure
- B. Promoting responsible innovation, competition, and collaboration will allow the US to lead in AI and unlock the technology's potential to solve some of society's most difficult challenges
- C. The responsible development and use of AI require a commitment to supporting US workers
- D. AI policies must be consistent with ... dedication to advancing equity and civil rights
- E. The interests of Americans who increasingly use, interact with, or purchase AI and AI-enabled products in their daily lives must be protected
- F. Americans' privacy and civil liberties must be protected as AI continues advancing
- G. It is important to manage the risks from the Federal Government's own use of AI and increase its internal capacity to regulate, govern, and support responsible use of AI to deliver better results for Americans
- H. The Federal Government should lead the way to global societal, economic, and technological progress, as the United States has in previous eras of disruptive innovation and change



## Section 3: Definitions ([link](#))

**AI** = “a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems use machine- and human-based inputs to perceive real and virtual environments; abstract such perceptions into models through analysis in an automated manner; and use model inference to formulate options for information or action.” ([Section 3\(b\)](#))

**AI model** = “a component of an information system that implements AI technology and uses computational, statistical, or machine-learning techniques to produce outputs from a given set of inputs.” ([Section 3\(c\)](#))

**Dual use foundation model** = “An AI model that is trained on broad data; generally uses self-supervision; contains at least tens of billions of parameters; is applicable across a wide range of contexts; and that exhibits, or could be easily modified to exhibit, **high levels of performance at tasks that pose a serious risk to security, national economic security, national public health or safety**, or any combination of those matters.... Models meet this definition even if they are provided to end users with technical safeguards that attempt to prevent users from taking advantage of the relevant unsafe capabilities.” ([Section 3\(k\)](#))

## Section 4: Ensuring the Safety and Security of AI Technology

- 4.1/ [Developing guidelines, standards and best practices for AI safety and security](#)
- 4.2/ [Ensuring Safe and Reliable AI](#)
- 4.3/ [Managing AI in Critical Infrastructure and in Cybersecurity](#)
- 4.4/ [Reducing risks at the intersection of AI and CBRN threats](#)
- 4.5/ [Reducing the risks posed by synthetic content](#)
- 4.6/ [Soliciting input on dual-use foundation models with widely available model weights](#)
- 4.7/ [Promoting safe release and preventing the malicious use of federal data for AI training](#)
- 4.8/ [Directing the development of a National security memorandum](#)

CBRN =  
chemical,  
biological,  
radiological,  
nuclear

By end July 2024 (270 days after publication) NIST should:

- Develop a variant of the **AI Risk Management Framework** tailored for generative AI **First draft (April 2024)**
- Extend Secure Software Development Framework to incorporate **secure development practices for generative AI and for dual-use foundation models** **First draft (April 2024)**
- Launch initiative to create guidance and **benchmarks for evaluating and auditing AI capabilities**, focusing on areas where AI could cause harm like cybersecurity and biosecurity
- Establish appropriate **guidelines for developers conducting AI red-teaming tests**

*Initial scope: Any model trained using **computing power > 10<sup>26</sup> flops** OR any model trained using primarily **biological sequence data and computing power > 10<sup>23</sup> flops** ([section 4.2\(b\)\(i\)](#))*

**By end Jan 2024** (90 days after publication) **companies developing or demonstrating intent to develop potential dual-use foundation models in scope need to provide:** ([section 4.2\(a\)\(i\)](#))

- **Details of ongoing/planned activities related to training, developing**, or producing dual-use foundation models, including the **physical and cybersecurity protections** taken to assure the integrity of that training process against sophisticated threats
- **Information about ownership and possession of the model weights**; and physical and cybersecurity measures taken to protect them
- **Results of performance in relevant AI red-team testing, and a description of any associated measures** taken to improve performance and strengthen overall model security. Prior to NIST guidance being developed, this should include results of any red-team testing relating to lowering the barrier to entry for the development, acquisition, and use of **biological weapons by non-state actors**; the discovery of **software vulnerabilities and development of associated exploits**; the use of software or tools to **influence real or virtual events**; the possibility for **self-replication or propagation**; and associated measures to meet safety objectives

*Initial scope: Any computing cluster that has a set of **machines physically co-located** in a single datacenter, **transitively connected by data center networking of over 100 Gbit/s**, and having a theoretical maximum computing **capacity of  $10^{20}$  flops** for training AI ([section 4.2\(b\)\(ii\)](#))*

**By end Jan 2024** (90 days after publication), any entity that acquires, develops or possesses a potential **large-scale computing cluster must report its location and the amount of total computing power** available in each cluster ([section 4.2\(a\)\(ii\)](#))

Additionally, plans to require US infrastructure providers (or resellers) to report ‘training runs’ by foreign persons of large AI models that could potentially be used in malicious cyber-activity, and verify their identity ([section 4.2\(c\)](#) and [4.2\(d\)](#))

### By end Jan 2024 (90 days after publication)

- **Assess potential risks related to AI in critical infrastructure sectors** and consider ways to mitigate vulnerabilities; repeat at least annually ([section 4.3\(a\)\(i\)](#))

### By end July 2024 (270 days after publication)

- Incorporate NIST's AI Risk Management Framework and other appropriate security guidance into **guidelines for critical infrastructure owners/operators** within 180 days ([section 4.3\(a\)\(iii\)](#)); then take steps to **mandate they are followed** via regulatory or other appropriate action ([section 4.3\(a\)\(iv\)](#))
- **Trial using AI capabilities to aid in discovery and remediation of vulnerabilities in critical US Government systems/networks** within 180 days ([section 4.3\(b\)\(ii\)](#)), and report on actions taken, vulnerabilities found and fixed, and lessons learned on how to deploy AI capabilities effectively for cyber defense ([section 4.3\(b\)\(iii\)](#))

# CBRN and synthetic content (not exhaustive)

4.4 Reducing risks at the intersection of AI and CBRN threats

4.5 Reducing the risks posed by synthetic content

## By end Feb 2024 (120 days after publication)

- Assess ways in which AI can increase biosecurity risks (e.g., generative AI models trained on biological data such as pathogens); and recommend mitigations ([section 4.4\(a\)\(ii\)](#))

## By end Apr 2024 (180 days after publication)

- Report to President on types of AI models that may present CBRN threats, including recommendations on training oversight, safety evaluation and guardrails ([section 4.4\(a\)\(i\)](#))

## By end Dec 2024 (420 days after publication)

- Issue guidance regarding the use of existing standards/methods/tools for authenticating, tracking provenance and watermarking synthetic content, as well as preventing generation of CSAM/non-consensual intimate imagery ([section 4.5\(a&b\)](#))

## Section 6: Supporting Workers ([link](#))

**By end Apr 2024** (180 days after publication) *(not exhaustive)*

- Submit reports to the President on the **labour market effects of AI, and how Federal programs could be used to respond** to future disruptions ([section 6\(a\)\(i & ii\)](#))
- Issue guidance to make clear that employers that deploy AI to monitor or **augment employees' work must continue to comply with protections that ensure workers are compensated** for their hours worked and other legal requirements ([section 6\(b\)\(ii\)](#))
- Prioritise **AI-related education** and related workforce development through existing programs ([section 6\(c\)](#))



## Section 10: Advancing Federal Government Use of AI

10.1/ [Providing Guidance for AI Management in Federal Government](#)

10.2/ [Increasing AI Talent in Government](#)

## By end Mar 2024 (150 days after publication)

- Issue guidance to Federal agencies on effective/appropriate use of AI, including required risk management practices, recommendations on testing/safeguards ([section 10.1\(b\)](#))

## By end Apr 2024 (180 days after publication)

- Develop guidance on the use of generative AI for work by the Federal government workforce ([section 10.1\(f\)\(iii\)](#))
- Facilitate access to Federal government-wide acquisition solutions for specified types of AI services and products (e.g., generative AI, specialised computing infrastructure) ([section 10.1\(h\)](#))

## By end May 2024 (210 days after publication)

- Designate a Chief AI Officer within each Federal agency ([section 10.1\(b\)\(i\)](#))
- Develop method to track/assess agencies ability to adopt AI, and manage its risks ([section 10.1\(c\)](#))



# 1. Recent milestones in AI regulation

2. Deeper dives into key legislation:

- **US Executive order** on safe, secure and trustworthy development and use of AI

- **European AI Act**

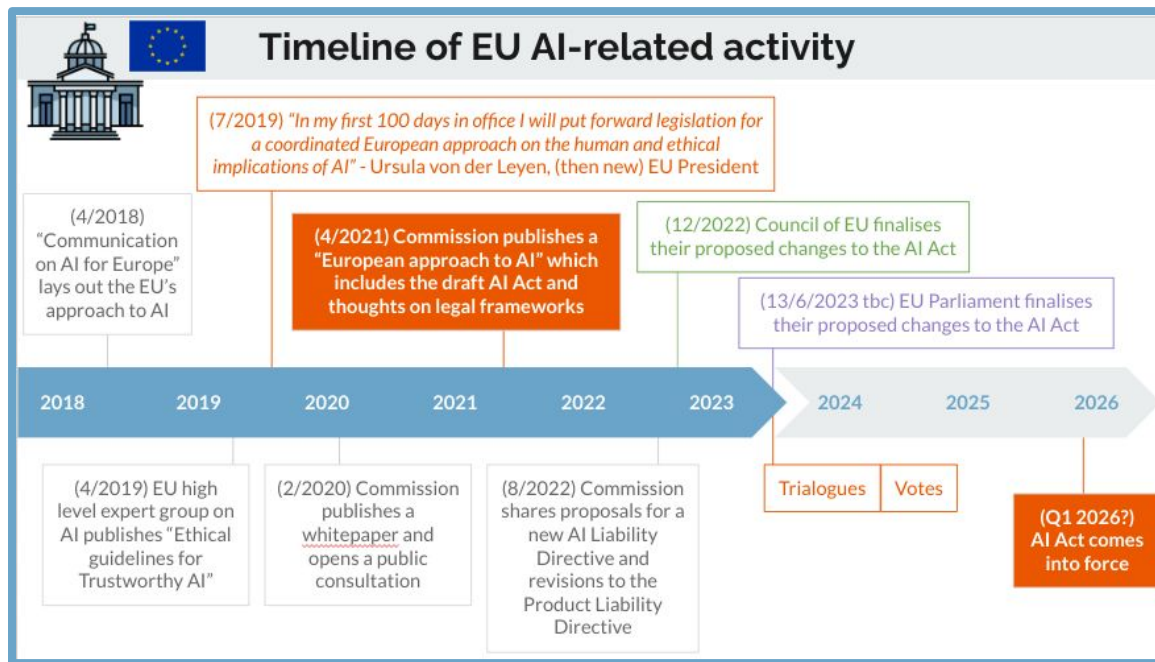


# European AI Act

## *Key points*

- Trialogues are over, and the text is now final – although still awaiting formal publication
- Broad shape of regulation is unchanged, although details have been finessed. Still waiting for clarification on standards

# The AI Act is finalised - but there's a delay



15/06/2023 Trialogues started  
08/12/2023 Trialogues ended

02/02/2024 EU Council vote  
13/03/2024 EU Parliament vote



**May/June 2024?: Publication**

... then "in force" 20 days later

BUT there is a built-in delay after it is "in force" before the Act gets applied:

- **Prohibited AI systems:** in 6 months (end 2024?)
- **General purpose AI models:** in 12 months (mid 2025?)
- **Remaining provisions on high risk AI:** in 2 years (mid 2026?)

---

# Scope

- Definition of AI
- Where and to whom it applies

# Definition of AI (Article 3(1))

European  
Commission  
(Apr 2021)

“An AI system means software that is developed with one or more of the techniques and approaches listed in Annex I and can for a given set of human-defined objectives generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with.”

## Annex I: AI techniques and approaches

- Machine learning approaches, including supervised, unsupervised and reinforcement learning, using deep learning, or a combination of such approaches;
- Logic- and knowledge-based approaches, including knowledge representation, (symbolic) reasoning and expert systems;
- Statistical approaches, Bayesian estimation, and probabilistic graphical models.

Council of EU  
(Dec 2020)

“An AI system means a machine-based system that is designed to operate with elements of autonomy to analyse and interpret data, to learn from human-provided data and inputs, infers how to achieve specific objectives using machine learning and/or logic- and knowledge-based approaches, and produces system-generated outputs such as content”

European  
Parliament  
(Jun? 2023)

“An AI system means a machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate outputs such as predictions, recommendations, or decisions, that influence physical or virtual environments”

Competing definitions as of mid-2023

# “AI system”

“**AI system** means a machine-based system designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment and that, for explicit or implicit objectives, **infers, from the input it receives, how to generate outputs such as predictions, content, recommendations or decisions that can influence physical or virtual environments**” (Article 3(1))

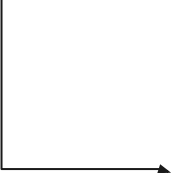
## Requirements for AI Act definition: (Recital 12)

- **Aligned with international organisations** to facilitate convergence and legal certainty
- **Flexibility** to accommodate rapid technological developments
- Based on key characteristics of AI that **distinguish it from simpler traditional software** systems or programming approaches. In particular, the capability to infer, using techniques such as machine learning, and logic/knowledge based approaches (but excluding systems based on rules defined solely by people to automatically execute operations)



# “General purpose” AI model/system

“**General purpose AI model** means an AI model, including when trained with a large amount of data using self-supervision at scale, that displays significant generality and is **capable to competently perform a wide range of distinct tasks** regardless of the way the model is placed on the market and that **can be integrated into a variety of downstream systems or applications**. This does not cover AI models that are used before release on the market for research, development and prototyping activities” ([Article 3\(63\)](#))



“**General purpose AI system** means an AI system which is **based on a general purpose AI model**, that has the **capability to serve a variety of purposes**, both for direct use as well as for integration in other AI systems ([Article 3\(66\)](#))

# Where and to whom the AI Act applies *(Article 2)*

**AI systems placed on the market or in service in the EU** regardless of provider's location  
[Council] + importers/distributors or manufacturers of products using an AI system

**AI systems where the output produced is used in the EU** regardless of provider's location  
[Parliament] + where the output is intended to be used in the EU

[Parliament] **AI systems by EU providers/distributors**

## Exclusions

- AI systems developed, produced, or provided by a natural or legal person acting in the course of their activity as a public authority or as an entity entrusted with a task of public interest or in the exercise of official authority
- AI systems used for law enforcement or judicial cooperation in criminal matters
- AI systems used solely for **scientific R&D**, or any **AI system R&D** for the purpose of testing (except in real world conditions) prior to be an AI system being put into service, provided it respects fundamental rights/EU law
- [Council] + AI systems used by individuals purely in a **personal non-professional capacity** (except for transparency requirements)
- [Parliament] + AI components (not including foundation models) provided under **free open source** licenses except if they are used as part of a high risk system or in a prohibited application

**In flux as of mid-2023**

# Clarification over “Providers” and “Deployers”

“**Provider** means a natural or legal person, public authority, agency or other body that **develops an AI system or a general purpose AI model OR that has an AI system or a general purpose AI model developed and places them on the market or puts the system into service under its own name or trademark, whether for payment or free of charge**” ([Article 3\(3\)](#))

“**Deployer** means any natural or legal person, public authority, agency or other body **using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity**” ([Article 3\(4\)](#))

# Where and to whom the AI Act applies ([Article 2\(1\)](#))

- **Providers** of AI systems or general-purpose AI models in the EU – regardless of where providers are located, or where systems/models were developed
- **Deployers** of AI systems which produce output that is used in the EU - regardless of the location of the deployer or AI system. (Includes also importers, distributors, manufacturers who rebadge/integrate AI systems into their own products)
- **Deployers** of AI systems who are located/established in the EU - regardless of where it is being used. (Except if used in course of a personal non professional activity)

## Exclusions

- High risk AI systems that relate to products covered by specific existing legislation ([Article 2\(2\)](#))
- AI systems that are exclusively for **military, defence or national security purposes**, regardless of the entity carrying out those activities ([Article 2\(3\)](#))
- AI systems used by **public authorities in a third country; or by an international organisation in the context of international cooperation or agreements for law/judicial enforcement** – BUT ONLY if there are adequate safeguards in place to protect fundamental rights/individual freedom ([Article 2\(4\)](#))
- AI systems and models (and their output) which are specifically developed/used for the **sole purpose of scientific research and development** ([Article 2\(6\)](#))
- Any **R&D** (excl. real world testing) prior to AI model/system being put on the market or in service ([Article 2\(8\)](#))
- AI systems released under **free and open source** licenses that are not part of a high risk AI system ([Article 2\(12\)](#))

# Other exemptions

- “Natural persons” deploying AI systems in the course of a **purely personal non-professional activity** do not need to meet deployer obligations ([Article 2\(10\)](#))
- **AI systems used in contexts classified as high risk, where it does not pose a significant risk of harm.** This will apply only if it does not involve profiling of people, and if it is intended only to do one or more of the following: ([Article 6\(3\)](#))
  - Perform a narrow procedural task
  - Improve the result of a previously completed human activity
  - Detect decision-making patterns or deviations from prior decision-making patterns and is not meant to replace or influence the previously completed human assessment, without proper human review
  - Perform a preparatory task for an assessment
- **Microenterprises providing high risk AI systems** will be able to comply with a simplified version of the quality management system ([Article 63](#))

---

# Restrictions

- General purpose AI models
- AI systems

---

# Restrictions

- **General purpose AI models**
- AI systems

# Two categories of general purpose AI model ([Article 51](#))

1

General purpose AI models with systemic risk

2

All other general purpose AI models



# Two categories of general purpose AI model ([Article 51](#))

1

General purpose AI models with systemic risk

General purpose AI model deemed to have “high impact capabilities”

[\(Article 51\(2\)\)](#) A general-purpose AI model shall be presumed to have high impact capabilities ... when the cumulative amount of compute used for its training measured in floating point operations (FLOPs) is greater than  $10^{25}$

[\(Article 51\(3\)\)](#) As technological developments advance (eg: algorithmic improvements, increased hardware efficiency), **the threshold will be amended when necessary** to reflect the state of the art, as well as supplemented with additional benchmarks and indicators

2

All other general purpose AI models

# Two categories of general purpose AI model ([Article 51](#))

1

General purpose AI models with systemic risk

General purpose AI model deemed to have “high impact capabilities”

Any general purpose AI model which the Commission decides has high capabilities or impact (e.g., following a qualified alert by their scientific panel).

2

All other general purpose AI models

In deciding this will take into account: ([Annex XIII](#))

- Number of parameters of the model
- Quality or size of the data set
- Amount of compute used for training the model (and related indicators such as training time/cost, energy used)
- Input and output modalities of the model; state-of-the-art thresholds for determining high-impact capabilities for each modality, and specific type of inputs and outputs (e.g. biological sequences);
- Benchmarks and evaluations of capabilities of the model, including considering the number of tasks without additional training, adaptability to learn new, distinct tasks, its degree of autonomy and scalability, the tools it has access to;
- Number of registered end-users; and availability. A high impact on the internal market shall be presumed when it has been made available to at least 10 000 registered EU business users

# How classification will work in practice ([Article 52](#))

- **Providers of general purpose AI models shall notify the Commission “without delay and in any event within 2 weeks”** after the requirements for being designated a model with systemic risk have been met – or it becomes known that they will be met.
  - If relevant: accompanying their notification, **providers can argue that their model should exceptionally NOT be classified as presenting systemic risk** due to its specific characteristics. The Commission will then decide if these arguments are strong enough to warrant an exception.
- **If the Commission learns of a model they’ve not been notified about that they believe meets the criteria, they will designate it to be of systemic risk.**
  - If relevant: Providers may offer a reasoned request to **reassess at the earliest six months after the designation decision**, by providing “objective, concrete and new reasons that have arisen”.

# Obligations for general purpose AI model providers ([Article 53](#))

- “Draw up, keep updated and make available” **documentation about the model**, with **extra details required for models deemed to pose systemic risk**.
  - Exemption: General purpose AI models which do not present systemic risk, that are accessible under a free/open license that “allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available.”
- Make publicly available a “**sufficiently detailed summary about the content used for training** of the general-purpose AI model”, and respect when copyright holders have requested that their works be excluded from data mining
- **Only for models deemed to pose systemic risk:**
  - Conduct and document state-of-the-art **adversarial testing**
  - Assess and **mitigate possible systemic risks** that may stem from model development/use
  - Track/document/report **serious incidents** and possible corrective measures without ‘undue delay’ to authorities
  - Ensure adequate **cybersecurity protection** for the model and associated physical infrastructure

# Obligations for general purpose AI model providers ([Article 53](#))

- “Draw up, keep updated and make available” **documentation about the model**, with **extra details required for models deemed to pose systemic risk**.
  - Exemption: General purpose AI models which do not present systemic risk, that are accessible under a free/open license that “allows for the access, usage, modification, and distribution of the model, and whose parameters, including the weights, the information on the model architecture, and the information on model usage, are made publicly available.”
- Make publicly available a “**sufficiently detailed summary about the content used for training** of the general-purpose AI model”, and respect when copyright holders have requested that their works be excluded from data mining
- **Only for models deemed to pose systemic risk:**
  - Conduct and document state-of-the-art **adversarial testing**
  - Assess and **mitigate possible systemic risks** that may stem from model development/use
  - Track/document/report **serious incidents** and possible corrective measures without ‘undue delay’ to authorities
  - Ensure adequate **cybersecurity protection** for the model and associated physical infrastructure

# Technical documentation for AI system providers

[Article 53 \(1b\)](#): Providers of general-purpose AI models need to make available up-to-date “information and documentation to providers of AI systems who intend to integrate the general-purpose AI model into their AI system.” **Documentation should convey a “good understanding of the capabilities and limitations of the general purpose AI model” and be sufficient to enable the AI system provider to “comply with their obligations”.**

To contain at a minimum (without compromising IP rights / trade secrets): ([Annex XII](#))

## *Basic details:*

- Architecture and number of parameters
- Tasks the model is intended to perform
- Modality (e.g., text, image, etc.) and format of the inputs and outputs and their maximum size (e.g., context window length)
- Information on the data used for training, testing and validation, where applicable, including type and provenance of data and curation methodologies

## *Usage specifications:*

- Applicable license; acceptable use policies
- Date of release; methods of distribution
- Type/nature of AI systems in which the model can be integrated; and technical means (e.g. infrastructure, tools, instructions) needed to do so
- How model can interact with hardware and software that is not part of the model itself
- Versions of relevant software related to the use of the general purpose AI model

# Technical documentation for authorities (1 of 2)

[Article 53 \(1a\)](#): Providers of general-purpose AI models need to make available up-to-date “**technical documentation of the model, including its training and testing process and the results of its evaluation**”, and provide it on request to the AI Office and national competent authorities

To contain: ([Annex XI Section 1](#))

## *Basic details and usage specifications:*

- Tasks the model is intended to perform
- Type/nature of AI systems in which the model can be integrated; and technical means (e.g. infrastructure, tools, instructions) needed to do so
- Applicable license; acceptable use policies
- Date of release; methods of distribution

## *Technical details related to model development:*

- **Design specifications including details of rationale and assumptions for key design choices**
  - Architecture and number of parameters
  - Modality, and format of inputs and outputs
  - Training methodologies/techniques
  - What the model is designed to optimise for and relevance of different parameters
- **Information on data used in training/testing/validation**
  - Type and provenance of data; curation methodologies (e.g. cleaning, filtering)
  - Number of data points, their scope and main characteristics
  - How data was obtained/selected; measures taken to detect unsuitable data sources or identifiable biases
- **Computational resources used** to train the model (e.g. number of FLOPs, training time) and other relevant details
- **Energy consumption** of the model (or estimate based on information about computational resources used).

# Technical documentation for authorities (2 of 2)

To contain... *continued* ([Annex XI Section 2](#))

*Additional information required for **general-purpose AI models with systemic risk**:*

- **Detailed description of the evaluation strategies, including evaluation results**, on the basis of available public evaluation protocols and tools or otherwise of other evaluation methodologies. Evaluation strategies shall include evaluation criteria, metrics and the methodology on the identification of limitations.
- Where applicable, **detailed description of the measures** put in place for the purpose of **conducting internal and/or external adversarial testing (e.g., red teaming), model adaptations, including alignment and fine-tuning**.
- Where applicable, **detailed description of the system architecture** explaining how software components build or feed into each other and integrate into the overall processing.



# What if you don't comply?

(Article 101): If the provider of a **general-purpose AI model intentionally/negligently does not comply: 15 million EUR fine, or up to 3%** of total worldwide annual turnover for the preceding financial year (whichever is higher)

Deadline for compliance: (Article 113)

- **Twelve months from the date of entry into force** – aka mid 2025?
- EXCEPT: penalties for general purpose AI model providers will not be imposed until 24 months from the date of entry into force

---

# Restrictions

- General purpose AI models
- **AI systems**

# AI Act takes a risk based approach

Minimal to low  
risk

Permitted with no restrictions

---

# AI Act takes a risk based approach

Minimal to low  
risk

Permitted with **no restrictions**

AI that needs  
transparency

Permitted if **transparency obligations** are met

- AI systems intended to interact with people
- Emotion recognition or biometric categorisation AI systems
- Generated text/image/audio/video

## Requirement to inform people they are interacting with AI

Providers of AI systems intended to directly interact with people must design/develop them in such a way that people are “**informed that they are interacting with an AI system, unless this is obvious** from the point of view of a natural person who is reasonably well-informed, observant and circumspect, taking into account the circumstances and the context of use.” ([Article 50\(1\)](#))

### Clarifications and exclusions ([Article 50\(1\)](#) and [Recital 132](#))

- Excludes AI **systems authorised by law to detect, prevent, investigate and prosecute** criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties, unless those systems are available for the public to report a criminal offence.
- In implementing this obligation, the characteristics of vulnerable individuals (e.g., elderly, disabled) should be taken into account, to the extent the system is intended to interact with them.

## Requirement to inform people exposed to biometric/emotion classification

Deployers shall **inform people who are exposed to the operation of emotion recognition or biometric categorisation systems**, and process data in accordance with existing regulations (e.g., GDPR) ([Article 50\(3\)](#))

### Clarifications and exclusions ([Article 50\(3\)](#) and [Recital 18](#))

- Excludes AI **systems permitted by law to detect, prevent, investigate and prosecute** criminal offences, subject to appropriate safeguards for the rights and freedoms of third parties.
- Emotion recognition systems are limited to those **that use biometric data** to infer emotions or intentions such as happiness, sadness, anger, surprise, disgust, embarrassment, excitement, shame, contempt, satisfaction and amusement.
- Excludes **detection of readily apparent expressions, gestures or movements** (e.g., facial expressions such as a frown or a smile, or hand gestures, or raised voice) **so long as they are not used to identify or infer emotion**.
- Excludes **detection of physical states such as pain or fatigue** (e.g., detecting fatigue in professional pilots or drivers for the purpose of preventing accidents)

## Requirement for generative AI output to be technically detectable

Providers of AI systems “generating synthetic audio, image, video or text content” must ensure the outputs are machine-readable and “**detectable as artificially generated or manipulated**”. They must also ensure that the technical solutions used to achieve this are “**effective, interoperable, robust and reliable as far as this is technically feasible**”, taking into account the context (e.g., content limitations, cost of implementation) and “the generally acknowledged state-of-the-art, as may be reflected in relevant technical standards.” ([Article 50\(2\)](#))

### Clarifications and exclusions ([Article 50\(2\)](#))

- Excludes AI systems performing an assistive function for standard editing, or that do not substantially alter the input data provided by the deployer or the semantics thereof.
- Excludes AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences.

## Requirement to disclose generative AI output

Deployers of AI systems that generate or manipulate **images, audio or video content constituting a deep fake, shall disclose it** has been artificially generated ([Article 50\(4\)](#))

### Clarifications and exclusions ([Article 50\(4\)](#))

- Where the content forms **part of an evidently artistic, creative, satirical, fictional** analogous work or programme, the transparency obligations set out in this paragraph are limited to disclosure of the existence of such generated or manipulated content **in an appropriate manner that does not hamper the display or enjoyment** of the work.
- Excludes **use authorised by law to detect, prevent, investigate and prosecute** criminal offence.

Deployers of an AI system that generates or manipulates **text published with the purpose of informing on matters of public interest shall disclose** that the content has been artificially generated. ([Article 50\(4\)](#))

### Clarifications and exclusions ([Article 50\(4\)](#))

- Excludes text that has undergone a process of **human review or editorial control, and where a natural or legal person holds editorial responsibility** for the publication of the content.
- Excludes **use authorised by law to detect, prevent, investigate and prosecute** criminal offence<sup>64</sup>



# What if you don't comply?

[\(Article 99\)](#): If you don't comply with the **obligations for transparency: 15 million EUR fine, or up to 3%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

Deadline for compliance: [\(Article 113\)](#)

- **24 months from the date of entry into force** – aka mid 2026?

# AI Act takes a risk based approach

Minimal to low  
risk

Permitted with **no restrictions**

AI that needs  
transparency

Permitted if transparency  
**obligations** are met

Unacceptable  
risk

**Prohibited**

- Manipulative or deceptive techniques
- Exploitation of vulnerable people
- Biometric categorisation systems
- Social scoring of people or groups
- Real-time remote biometric identification in public areas
- Profiling to predict risk of committing an offense
- Untargeted scraping of facial images
- Inferring emotions in workplaces and educational institutions (except for medical/safety reasons)

## *Prohibitions on using AI systems to manipulate*

**Subliminal (or) purposefully manipulative/deceptive techniques** that materially distort behaviour, by appreciably impairing a person's ability to make an informed decision, causing them to take a decision they would not otherwise have taken that is likely to cause significant harm ([Article 5\(1a\)](#))

**Exploitation of people's vulnerabilities** due to their age, disability, or specific social/economic situation that materially distorts their behaviour in a manner that is reasonably likely to cause significant harm ([Article 5\(1b\)](#))

### Clarifications and exclusions ([Recital 29](#))

- It is **not necessary for the provider to have the intention to cause significant harm**, as long as such harm results from the manipulative or exploitative AI-enabled practices
- **Excludes lawful practices in context of medical treatment** when carried out in accordance with applicable legislation (e.g., explicit consent of individuals)
- **Excludes common and legitimate commercial practices (e.g., in advertising)** that are in compliance with applicable laws

## Prohibitions on using AI systems for profiling

**Biometric categorisation systems** used to “deduce or infer (someone’s) race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation” (exclusion: use in law enforcement) ([Article 5\(1g\)](#))

**Social scoring** of people or groups “based on their social behaviour or known, inferred or predicted personal or personality characteristics”, when it leads to detrimental or unfavourable treatment in social contexts unrelated to the contexts in which the data was originally collected, or that is unjustified or disproportionate to their social behaviour or its gravity ([Article 5\(1c\)](#))

### Clarifications and exclusions ([Recitals 15](#) and [16](#))

- Systems used **solely to verify identity to grant access** to services/devices/premises
- Systems that are a purely **ancillary feature to an allowed functionality** (e.g., using facial/body filters so a person can preview themselves wearing a product to aid in purchase decision; or to support adding/modifying images on social networks)

## Prohibitions on using AI systems in policing

### **The use of real-time remote biometric identification systems in publicly accessible spaces** for purposes of law enforcement ([Article 5\(1h\)](#))

#### Clarifications and exclusions ([Recital 17](#); [Articles 5\(1h\)](#) and [5\(2\)](#))

- Rule on 'real time' use cannot be circumvented by providing for minor delays
- Excludes biometric ID systems used in **targeted searches for victims** of abduction, trafficking, sexual exploitation and missing persons; or to help prevent a substantial and **imminent terrorist attack threat**; or as part of investigation/prosecution of a suspected criminal in relation to an **offence with maximum custodial sentence of 4+ years**
- Law enforcement use should comply with safeguards imposed by national legislations, including a 'fundamental human rights impact assessment'. Each use should receive prior judicial authorisation (except in emergencies)

### Using AI to **predict the risk of a person committing a criminal offence, based solely on profiling** or assessing their personality traits or characteristics ([Article 5\(1d\)](#))

#### Clarifications and exclusions ([Article 5\(1d\)](#))

- Excludes using AI to **support human assessment of the involvement of a person** in a criminal activity, which is already based on objective and verifiable facts directly linking them to it

## Prohibitions on using AI systems to expand surveillance

Making available or using AI systems specifically to “**create or expand facial recognition databases through the untargeted scraping of facial images** from the internet or CCTV footage” ([Article 5\(1e\)](#))

Making available or using AI systems specifically to “**infer emotions of a natural person in... workplaces and educational institutions,**” except where it is intended for medical or safety reasons ([Article 5\(1f\)](#))

### Clarifications and exclusions ([Recital 18](#))

- Limited to AI **systems that use biometric data** to infer emotions or intentions such as happiness, sadness, anger, surprise, disgust, embarrassment, excitement, shame, contempt, satisfaction and amusement.
- Excludes **detection of readily apparent expressions, gestures or movements** (e.g., facial expressions such as a frown or a smile, or hand gestures, or raised voice) **so long as they are not used to identify or infer emotion.**
- Excludes **detection of physical states such as pain or fatigue** (e.g., detecting fatigue in professional pilots or drivers for the purpose of preventing accidents)

# What if you don't comply?

[\(Article 99\)](#): If you engage in any of the **prohibited AI practices: 35 million EUR fine, or up to 7%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

## Deadline for compliance:

- **Six months from the date of entry into force** – aka end 2024? ([Article 113](#))
- EXCEPT: Any systems that are components of specified large IT systems related to border and migration (e.g., Schengen Information System; full list in [Annex 10](#)). So long as they were already on the market or put into service at least 36 months prior to the AI Act coming into force, they have until end 2030 to comply ([Article 111](#))

# AI Act takes a risk based approach

Minimal to low  
risk

Permitted with **no restrictions**

AI that needs  
transparency

Permitted if **transparency obligations** are met

High risk

Permitted **subject to compliance** with AI Act requirements and ex-ante conformity assessment

*Specified applications in key fields:*

Regulated products

Biometrics

Critical infrastructure

Education

Essential services

Employment

Law enforcement

Legal

Border control

Unacceptable  
risk

**Prohibited**



# Classification of high risk AI systems... 1 of 3 ([Article 6](#) and [Annex 3](#))

## Regulated products

AI systems that are a product already subject to third party certification; or being used as safety components in such products (e.g., toys, elevators, motor vehicles, gas-burning appliances, medical devices, etc)

---

## Biometrics

AI systems used for non-prohibited remote biometric identification (except solely to confirm ID), categorisation of people according to sensitive attributes; or emotion recognition

---

## Critical infrastructure

AI systems used as a safety component in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity

Education  
Employment  
Essential services  
Law enforcement  
Border control  
Legal

# Classification of high risk AI systems... 2 of 3 ([Annex 3](#))

Regulated products  
Biometrics  
Critical infrastructure

## Education

AI systems used to determine people's access or admission to study; evaluate learning outcomes (including to steer learning process); and to monitor/detect prohibited behaviour

## Employment

AI systems used in recruitment (e.g., to place targeted job ads, analyse and filter job applications, evaluate candidates); and to make decisions relating to work-related relationships (e.g., promotion, termination, task allocation, performance evaluation)

## Essential services

Using to assess creditworthiness (except to detect fraud), determine eligibility for public assistance (including healthcare); prioritise dispatch of emergency services; assess risk and set pricing for health and life insurance

Law enforcement  
Border control  
Legal

# Classification of high risk AI systems... 3 of 3 ([Annex 3](#))

Regulated products  
Biometrics  
Critical infrastructure  
Education  
Employment  
Essential services

**Law  
enforcement**

AI systems used to assess risk of a person becoming a crime victim; in tools like polygraphs; to assess the risk of offending or to profile people and personality traits; and to evaluate the reliability of evidence in the course of a criminal investigation/prosecution

**Border control**

AI systems used in tools like polygraphs; to assess entry risk (e.g., security, health, irregular migration); to assess asylum/visa/residence claims; or to detect/recognise people except to verify travel documents

**Legal**

AI systems used to assist a judicial authority to research/interpret facts or in applying laws; or to influence the outcome of an election including voting behaviour (except tools that people are not directly exposed to, eg: tools used to organise political campaigns)

# Requirements for high risk AI systems

Processes

**Risk management system** ([Article 9](#)) - Identify, analyse and evaluate known/foreseeable risks, and adopt suitable risk management measures, including testing

**Data governance** ([Article 10](#)) - Data sets shall be relevant, representative, free of errors and complete, taking into account the intended purpose/context for AI system use

**Transparency** ([Article 13](#)) - Provide users with clear/concise information on the system's intended purpose, capabilities and limitations, and maintenance requirements

**Human oversight** ([Article 14](#)) - Monitoring to allow early detection of issues; measures to facilitate interpretability of output, and allow to override/disregard when appropriate

**Accuracy, robustness & cybersecurity** ([Article 15](#)) - Resilient to errors, inconsistencies, attacks; backup fail-safe plans

**Quality management system** ([Article 17](#)) - Strategy for regulatory compliance, including techniques/procedures/actions; Monitoring system and procedure to report serious errors

**Documentation & record keeping** ([Articles 11 & 12](#)) - Provide authorities with necessary information for them to assess compliance

- Intended purpose, capabilities and performance limitations
- Design specifications ('logic', assumptions, data used, standards applied, interpretability)
- Automatic recording of events ('logs') while the AI system is operating

Design specifications

# CEN/CENELEC is developing European standards (due 04/2025)

1. Risk management systems for AI systems

**Risk management system** ([Article 9](#)) - Identify, analyse and evaluate known/foreseeable risks, and adopt suitable risk management measures, including testing

**Data governance** ([Article 10](#)) - Data sets shall be relevant, representative, free of errors and complete, taking into account the intended purpose/context for AI system use

**Transparency** ([Article 13](#)) - Provide users with clear/concise information on the system's intended purpose, capabilities and limitations, and maintenance requirements

**Human oversight** ([Article 14](#)) - Monitoring to allow early detection of issues; measures to facilitate interpretability of output, and allow to override/disregard when appropriate

**Accuracy, robustness & cybersecurity** ([Article 15](#)) - Resilient to errors, inconsistencies, attacks; backup fail-safe plans

**Quality management system**

([Article 17](#)) - Strategy for regulatory compliance, including techniques/procedures. Monitoring system and report serious errors

**Documentation & record keeping** ([Articles 11 & 12](#)) - Provide authorities with necessary information for them to assess compliance

- Intended purpose, performance
- Design assumptions applied, interpretability
- Automatic recording of events ('logs') while the AI system is operating

2. Governance and quality of datasets used to build AI systems

4. Transparency and information provisions for users of AI systems

5. Human oversight of AI systems

6. Accuracy specifications for AI systems

7. Robustness specifications for AI systems

8. Cybersecurity specifications for AI systems

9. Quality management systems for providers of AI systems, including post-market monitoring processes

3. Record keeping through logging capabilities by AI systems

10. Conformity assessment for AI systems

# Who is responsible? (1 of 3)

## Providers [\(Article 16\)](#)

## Deployers [\(Article 26\)](#)

**Data  
governance**

**Transparency**

**Human  
oversight**

**Accuracy,  
robustness,  
cybersecurity**

**Ensure the AI system undergoes  
the relevant conformity  
assessment procedure prior to  
being put into service**

Draw up EU declaration of  
conformity and affix CE label

Upon the request of authorities,  
demonstrate conformity with  
requirements, and provide access  
to automated logs [\(Article 21\)](#)

To the extent that exercise control:  
ensure inputs are relevant to  
intended purposes and sufficiently  
representative

Provide additional transparency  
for certain kinds of high risk AI  
systems (e.g., alerting workers when  
deployed in workplaces)

Assign human oversight to  
people who have the necessary  
skills/authority

# Who is responsible? (2 of 3)

## Providers

## Deployers ([Article 26](#))

Risk management system

Identify/analyse foreseeable risks and adopt suitable risk management measures ([Article 9](#))

Test to ensure the system performs consistently for its intended purpose

Carry out a data protection impact assessment

Quality management system

Put in place and document a quality management system ([Article 17](#))

Inform authorities of any non-compliance and corrective actions taken ([Article 20](#))

Use and monitor the system in line with instructions  
Report serious incidents/malfunctioning

Documentation & record keeping

Draw up technical documentation ([Article 18](#))

Keep automatically generated logs for at least 6 months (if in their control) ([Article 19](#))

Keep automatically generated logs for at least 6 months (if in their control)

Processes

# Who is responsible? (3 of 3)

## Importers of high risk AI systems

[\(Article 23\)](#)

Before placing on the market, verify that:

- Relevant conformity assessment has been carried out by the provider
- Provider has drawn up the required technical documentation
- CE marking is affixed and accompanied by conformity declaration

Cooperate with authorities as required

## Distributors of high risk AI systems

[\(Article 24\)](#)

Before making available:

- Verify the CE marking is affixed and accompanied by conformity declaration
- Ensure provider and importer (as applicable) have complied their obligations. Do not make available if have reason to consider they've not

Ensure storage/transport conditions do not jeopardise compliance of system

Cooperate with authorities as required



# Who is responsible? (3 of 3)

## Importers of high risk AI systems

(Article 23)

Before placing on the market, verify that:

- Relevant conformity assessment has been carried out by the provider
- Provider has drawn up technical documentation
- CE marking is accompanied by a declaration

Cooperate with authorities

## Contributors of high risk AI systems

(Article 24)

able:

**(Article 25) A third party takes over the responsibilities of the provider if:**

- they put their name/trademark on it (unless otherwise agreed);
- if they make a substantial modification to the system or to its intended purpose

Cooperate with authorities as required

# What if you don't comply?

([Article 99](#)): If you don't comply with the **obligations for high risk AI systems: 15 million EUR fine, or up to 3%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

If you supply incorrect, incomplete or **misleading information to authorities in reply to a request: 7.5 million EUR fine, or up to 1%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

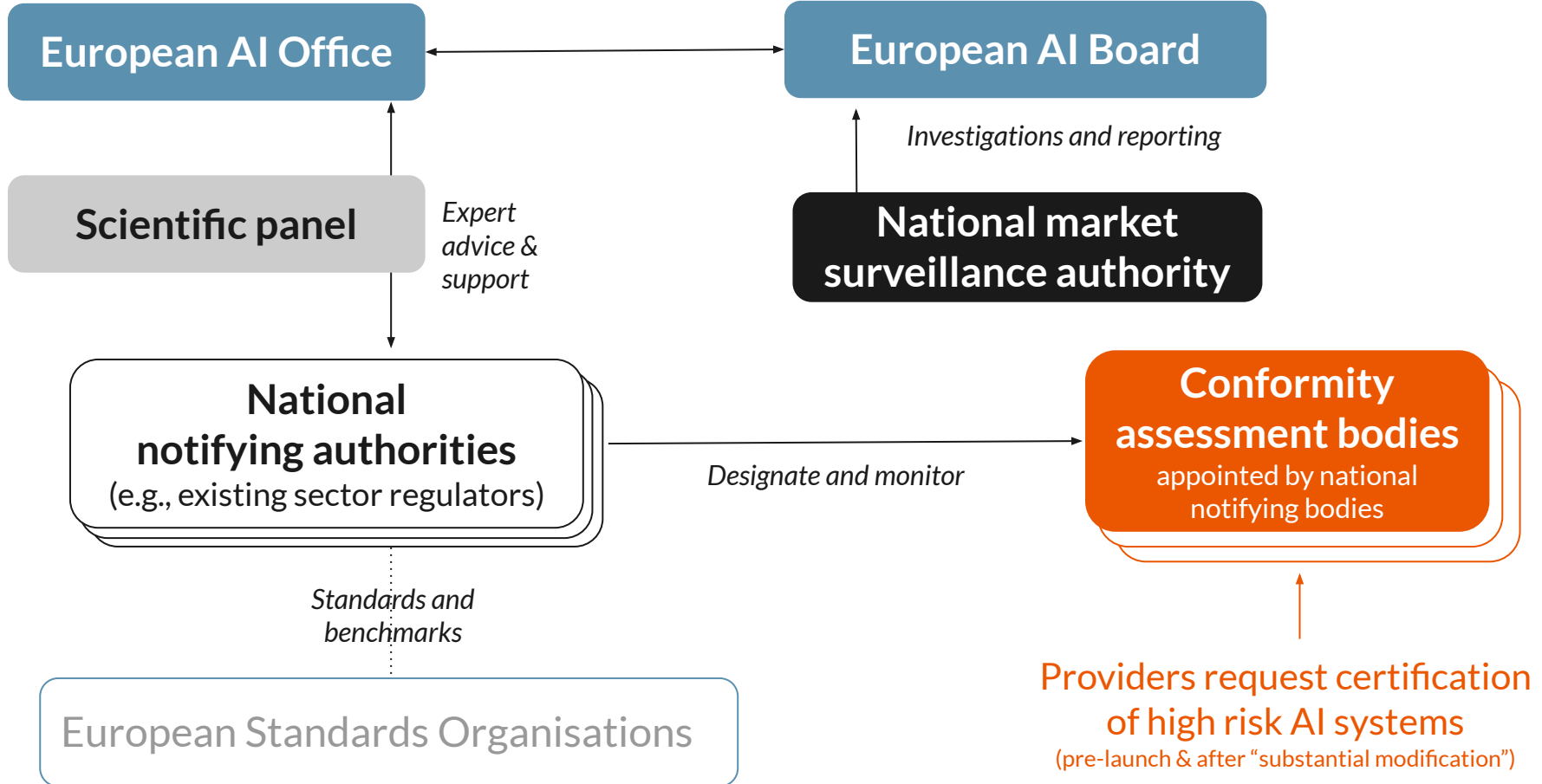
Deadline for compliance: ([Article 113](#))

- **24 months from the date of entry into force** – aka mid 2026?
- EXCEPT: High risk AI systems used in or as safety components, which have 36 months to comply from the data of entry into force

---

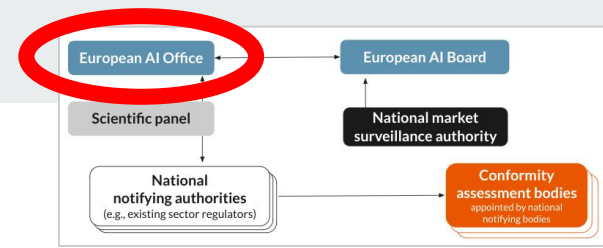
# Enforcement

# How the AI Act will be enforced



# Introduction of the “European AI Office”

([Article 64](#); [link](#)) Goal is to be the European Commission’s “centre of AI expertise” and the “foundation for a single European AI governance system”. Part of the Directorate-General for Communication Networks, Content and Technology.

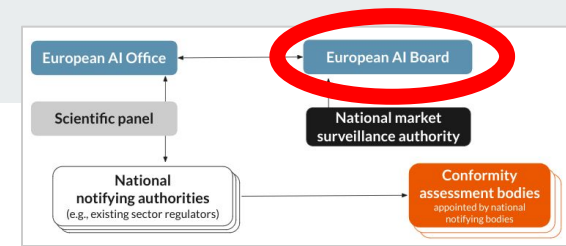


The AI Office’s role and powers include:

- **Enforcing obligations on general purpose AI model providers** ([Article 88](#))
- **Ongoing monitoring and evaluation of general purpose AI models** or appointing independent experts to do so ([Article 89](#)). This includes **power to request access to the model through APIs or other appropriate means, incl. source code** ([Article 92](#))
- Monitoring and supervision of general purpose AI systems where the provider of the system is the same as the model provider ([Article 75](#))
- Requesting general purpose AI model providers take measures to mitigate issues identified; or even to restrict/withdraw its availability ([Article 93](#))
- **Developing guidelines on practical implementation of the AI Act** ([Article 96](#))

# Introduction of the “European AI Board”

**(Article 65)** Composed of one representative per Member State, who must have relevant competencies and authority to contribute actively to the Board’s tasks. The **European Data Protection Supervisor shall participate as observer**. The **AI Office shall also attend the Board’s meetings, without taking part in the votes**. Other national and Union authorities, bodies or experts may be invited to the meetings by the Board on a case by case basis, where relevant issues are discussed.

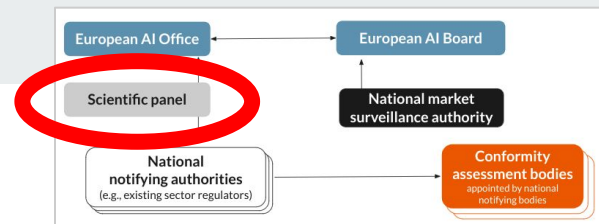


**(Article 66)** European AI Board’s role is to advise and assist the Commission and Member states on consistent and effective application of the AI Act. Example tasks:

- Help with coordination between different national authorities
- Collect/share technical and regulatory best practices
- Advise on the enforcement of rules on general purpose AI models
- Issue recommendations/opinions on matters relevant to the AI Act’s implementation (e.g., trends in AI competitiveness, trends in AI value chains)
- Assist national authorities/AI office to develop organisational and technical expertise

# Introduction of the “Scientific Panel”

([Article 68](#)) The scientific panel shall consist of **experts selected by the Commission on the basis of up-to-date scientific or technical expertise** in the field of AI. **All experts must be independent** from any provider of AI systems or general purpose AI models



## Tasks of the Scientific Panel:

- **Provide a qualified alert** if they have reasons to suspect a general purpose AI system poses concrete identifiable risk; or if they believe an AI system is high risk without having been appropriately designated ([Article 90](#))
- Members of the panel may be **invited to conduct evaluations** of general purpose AI models by the AI Office ([Article 92](#))
- Provide advice/support to Member States on enforcement on request, possibly subject to pre-agreed fees ([Article 69](#))

# Introduction of “National Competent Authorities”

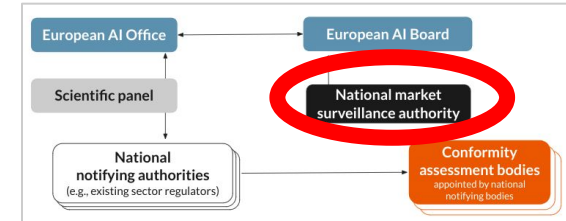
## Notifying authority ([Article 28](#))

- Each Member state to establish or designate at least one notifying authority and ensure it has adequate competent personnel
- Responsible for assessment/designation/monitoring of conformity assessment bodies



## Market surveillance authority ([Article 70](#))

- Each member state to establish or designate at least one market surveillance authority, and ensure they are provided with adequate technical/financial/human resources





# Testing arrangements

## AI regulatory sandboxes:

- Member states to establish at least one sandbox at the national level, within 24 months of the AI Act entering into force ([Article 57](#))
- Commission will provide common operating principles for the sandbox, including criteria for participation, monitoring, exiting, etc ([Article 58](#))

## Real world testing: ([Article 60](#))

- A real-world testing plan must be submitted for approval to the ‘market surveillance authority’ in the Member State where testing is taking place, prior to testing commencing. If there is no response within 30 days, it can be understood to have been approved
- Testing cannot last for longer than is necessary for its objectives; and no longer than 6 months (with possibility of a 6 month extension on request)
- Participants in testing must give their informed consent ([Article 61](#))

## What if you don't comply? (Articles [99](#) and [101](#))

If you engage in any of the **prohibited AI practices: 35 million EUR fine, or up to 7%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

If you don't comply with the **obligations for high risk AI systems, or obligations for transparency: 15 million EUR fine, or up to 3%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

If you supply incorrect, incomplete or **misleading information to authorities in reply to a request: 7.5 million EUR fine, or up to 1%** of total worldwide annual turnover for the preceding financial year (whichever is higher for large companies; whichever is lower for SMEs)

If the provider of a **general-purpose AI model intentionally/negligently does not comply with documentation requirements: 15 million EUR fine, or up to 3%** of total worldwide annual turnover for the preceding financial year (whichever is higher)



# In a nutshell...



**Increased focus on 'foundation models' and broader AI safety issues** beyond fairness/data – driven initially by UK

The **US is no longer on the sidelines**, and is leading by example with AI regulation for Federal agencies

>> Executive Order (Oct 2024)

AI oversight in **UK and US is predominantly sector-specific**, Vs. **Europe has taken a broader 'horizontal' approach**

>> AI Act (May/June? 2024) – but delayed enforcement

---

**Thank you**