AI Act
Consolidated text

TL;DR

The AI Act classifies AI according to its risk:

- Unacceptable risk is prohibited (e.g. social scoring systems and manipulative AI).
- Most of the text addresses high-risk AI systems, which are regulated.
- A smaller section handles limited risk AI systems, subject to lighter transparency obligations: developers and deployers must ensure that end-users are aware that they are interacting with AI (chatbots and deepfakes).
- Minimal risk is unregulated (including the majority of AI applications currently available on the EU single market, such as AI enabled video games and spam filters – at least in 2021; this is changing with generative AI).

The majority of obligations fall on providers (developers) of high-risk AI systems.

- Those that intend to place on the market or put into service high-risk AI systems in the EU, regardless of whether they are based in the EU or a third country.
- And also third country providers where the high risk AI system's output is used in the EU.

Deployers are natural or legal persons that deploy an AI system in a professional capacity, not affected end-users.

- Deployers of high-risk AI systems have some obligations, though less than providers (developers).
- This applies to deployers located in the EU, and third country users where the AI system's output is used in the EU.

General purpose AI (GPAI):

- All GPAI model providers must provide technical documentation, instructions for use, comply with the Copyright Directive, and publish a summary about the content used for training.
- Free and open licence GPAI model providers only need to comply with copyright and publish the training data summary, unless they present a systemic risk.
- All providers of GPAI models that present a systemic risk – open or closed – must also conduct model evaluations, adversarial testing, track and report serious incidents and ensure cybersecurity protections.

Prohibited AI systems (Chapter II, Art. 5)

AI systems:

- deploying **subliminal, manipulative, or deceptive techniques** to distort behaviour and impair informed decision-making, causing significant harm.
- **exploiting vulnerabilities** related to age, disability, or socio-economic circumstances to distort behaviour, causing significant harm.
- **social scoring**, i.e., evaluating or classifying individuals or groups based on social behaviour or personal traits, causing detrimental or unfavourable treatment of those people.
- **assessing the risk of an individual committing criminal offenses** solely based on profiling or personality traits, except when used to augment human assessments based on objective, verifiable facts directly linked to criminal activity.
- **compiling facial recognition databases** by untargeted scraping of facial images from the internet or CCTV footage.
- **inferring emotions in workplaces or educational institutions**, except for medical or safety reasons.
- **biometric categorisation systems** inferring sensitive attributes (race, political opinions, trade union membership, religious or philosophical beliefs, sex life, or sexual orientation), except labelling or filtering of lawfully acquired biometric datasets or when law enforcement categorises biometric data.
- **'real-time' remote biometric identification (RBI) in publicly accessible spaces for law enforcement**, except when:
        - targeted searching for missing persons, abduction victims, and people who have been human trafficked or sexually exploited;
        - preventing specific, substantial and imminent threat to life or physical safety, or foreseeable terrorist attack; or
        - identifying suspects in serious crimes (e.g., murder, rape, armed robbery, narcotic and illegal weapons trafficking, organised crime, and environmental crime, etc.).
    - Using AI-enabled real-time RBI is only allowed **when not using the tool would cause harm,** particularly regarding the seriousness, probability and scale of such harm, and must account for affected persons' rights and freedoms.
    - Before deployment, police must complete a **fundamental rights impact assessment** and **register the system in the EU database**, though, in duly justified cases of urgency, deployment can commence without registration, provided that it is registered later without undue delay.
    - Before deployment, they also must obtain **authorisation from a judicial authority or independent administrative authority**[1], though, in duly justified cases of urgency, deployment can commence without authorisation, provided that authorisation is requested within 24 hours. If authorisation is rejected, deployment must cease immediately, deleting all data, results, and outputs.

[1] Independent administrative authorities may be subject to greater political influence than judicial authorities (Hacker, 2024).

High risk AI systems (Chapter III)

Classification rules for high-risk AI systems (Art. 6)

High risk AI systems are those:

- used as a safety component or a product covered by EU laws in Annex I **AND** required to undergo a third-party conformity assessment under those Annex I laws; **OR**
- those under Annex III use cases (below), except if:
  - the AI system performs a narrow procedural task;
  - improves the result of a previously completed human activity;
  - detects 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; or
  - performs a preparatory task to an assessment relevant for the purpose of the use cases listed in Annex III.
- The Commission can add or modify the above conditions through delegated acts if there is concrete evidence that an AI system falling under Annex III does not pose a significant risk to health, safety and fundamental rights. They can also delete any of the conditions if there is concrete evidence that this is needed to protect people.
- AI systems are always considered high-risk if it profiles individuals, i.e. automated processing of personal data to assess various aspects of a person's life, such as work performance, economic situation, health, preferences, interests, reliability, behaviour, location or movement.
- Providers that believe their AI system, which fails under Annex III, is not high-risk, must document such an assessment before placing it on the market or putting it into service.
- 18 months after entry into force, the Commission will provide guidance on determining if an AI system is high risk, with list of practical examples of high-risk and non-high risk use cases.

Requirements for providers of high-risk AI systems (Art. 8-17)

High risk AI providers must:

- Establish a **risk management system** throughout the high risk AI system's lifecycle;
- Conduct **data governance**, ensuring that training, validation and testing datasets are relevant, sufficiently representative and, to the best extent possible, free of errors and complete according to the intended purpose.
- Draw up **technical documentation** to demonstrate compliance and provide authorities with the information to assess that compliance.
- Design their high risk AI system for **record-keeping** to enable it to automatically record events relevant for identifying national level risks and substantial modifications throughout the system's lifecycle.
- Provide **instructions for use** to downstream deployers to enable the latter's compliance.
- Design their high risk AI system to allow deployers to implement **human oversight.**
- Design their high risk AI system to achieve appropriate levels of **accuracy, robustness, and cybersecurity**.
- Establish a **quality management system** to ensure compliance.

Annex III use cases
Non-banned biometrics: ▪ Remote biometric identification systems, excluding biometric verification that confirm a person is who they claim to be. ▪ Biometric categorisation systems inferring sensitive or protected attributes or characteristics. ▪ Emotion recognition systems.
Critical infrastructure: ▪ Safety components in the management and operation of critical digital infrastructure, road traffic and the supply of water, gas, heating and electricity.
Education and vocational training: ▪ AI systems determining access, admission or assignment to educational and vocational training institutions at all levels. ▪ Evaluating learning outcomes, including those used to steer the student's learning process. ▪ Assessing the appropriate level of education for an individual. ▪ Monitoring and detecting prohibited student behaviour during tests.
Employment, workers management and access to self-employment: ▪ AI systems used for recruitment or selection, particularly targeted job ads, analysing and filtering applications, and evaluating candidates. ▪ Promotion and termination of contracts, allocating tasks based on personality traits or characteristics and behaviour, and monitoring and evaluating performance.
Access to and enjoyment of essential public and private services: ▪ AI systems used by public authorities for assessing eligibility to benefits and services, including their allocation, reduction, revocation, or recovery. ▪ Evaluating creditworthiness, except when detecting financial fraud. ▪ Evaluating and classifying emergency calls, including dispatch prioritising of police, firefighters, medical aid and urgent patient triage services. ▪ Risk assessments and pricing in health and life insurance.
Law enforcement: ▪ AI systems used to assess an individual's risk of becoming a crime victim. ▪ Polygraphs. ▪ Evaluating evidence reliability during criminal investigations or prosecutions. ▪ Assessing an individual's risk of offending or re-offending not solely based on profiling or assessing personality traits or past criminal behaviour. ▪ Profiling during criminal detections, investigations or prosecutions.
Migration, asylum and border control management: ▪ Polygraphs. ▪ Assessments of irregular migration or health risks. ▪ Examination of applications for asylum, visa and residence permits, and associated complaints related to eligibility. ▪ Detecting, recognising or identifying individuals, except verifying travel documents.
Administration of justice and democratic processes: ▪ AI systems used in researching and interpreting facts and applying the law to concrete facts or used in alternative dispute resolution.

- Influencing elections and referenda outcomes or voting behaviour, excluding outputs that do not directly interact with people, like tools used to organise, optimise and structure political campaigns.

General purpose AI (GPAI) (Chapter V)

GPAI 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.

GPAI 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.

GPAI systems may be used as high risk AI systems or integrated into them. GPAI system providers should cooperate with such high risk AI system providers to enable the latter's compliance.

All providers of GPAI models must (Art. 53):

- Draw up **technical documentation**, including training and testing process and evaluation results.
- Draw up **information and documentation to supply to downstream providers** that intend to integrate the GPAI model into their own AI system in order that the latter understands capabilities and limitations and is enabled to comply.
- Establish a policy to **respect the Copyright Directive.**
- Publish a **sufficiently detailed summary about the content used for training** the GPAI model.

Free and open licence GPAI models – whose parameters, including weights, model architecture and model usage are publicly available, allowing for access, usage, modification and distribution of the model – only have to comply with the latter two obligations above, unless the free and open licence GPAI model is systemic.

GPAI models are considered systemic when the cumulative amount of compute used for its training is greater than 10^25 floating point operations per second (FLOPS) (Art. 51). Providers must notify the Commission if their model meets this criterion within 2 weeks (Art. 52). The provider may present arguments that, despite meeting the criteria, their model does not present systemic risks. The Commission may decide on its own, or via a qualified alert from the scientific panel of independent experts, that a model has high impact capabilities, rendering it systemic.

In addition to the four obligations above, providers of GPAI models with systemic risk must also (Art. 55):

- Perform **model evaluations**, including conducting and documenting **adversarial testing** to identify and mitigate systemic risk.
- **Assess and mitigate possible systemic risks**, including their sources.
- **Track, document and report serious incidents** and possible corrective measures to the AI Office and relevant national competent authorities without undue delay.
- Ensure an adequate level of **cybersecurity protection**.

All GPAI model providers may demonstrate compliance with their obligations if they voluntarily adhere to codes of practice until European harmonised standards are published, compliance with which will lead to a presumption of conformity (Art. 56). Providers that don't adhere to codes of practice must demonstrate **alternative adequate means of compliance** for Commission approval.

Codes of practice (Art. 56)

- Will account for international approaches.
- Will cover but not necessarily limited to the above obligations, particularly the relevant information to include in technical documentation for authorities and downstream providers, identification of the type and nature of systemic risks and their sources, and the modalities of risk management accounting for specific challenges in addressing risks due to the way they may emerge and materialise throughout the value chain.
- AI Office may invite GPAI model providers, relevant national competent authorities to participate in drawing up the codes, while civil society, industry, academia, downstream providers and independent experts may support the process.

Governance (Chapter VI)

- The AI Office will be established, sitting within the Commission, to monitor the effective implementation and compliance of GPAI model providers (Art. 64).
- Downstream providers can lodge a complaint regarding the upstream providers infringement to the AI Office (Art. 89).
- The AI Office may conduct evaluations of the GPAI model to (Art. 92):
    - assess compliance where the information gathered under its powers to request information is insufficient.
    - Investigate systemic risks, particularly following a qualified report from the scientific panel of independent experts (Art. 90).

Timelines

- After entry into force, the AI Act will apply by the following deadlines:
    - 6 months for prohibited AI systems.
    - 12 months for GPAI.
    - 24 months for high risk AI systems under Annex III.
    - 36 months for high risk AI systems under Annex I.
- Codes of practice must be ready 9 months after entry into force.