GOLDEN NUGGETS

INSIGHT SERIES 138

PART 1: THE EU AI ACT IN OUTLINE
PART 2: THE EU AI ACT'S CODES OF PRACTICE

PART II

EU AI ACT

THE CODES OF PRACTICE

12TH JULY 2025

ADLSOLICITORS.COM IN ASSOCIATION WITH

CENTRE FOR THE ASSESSMENT OF ARTIFICIAL INTELLIGENCE RISKS AND OPPORTUNITIES
CAAIRO.COM
(SEE PODCASTS AT CAAIRO.COM)

DR NICK LOCKETT

RIGHTS:

No liability for anny content or reliance upon content is accepted by the authors. Pleasae see legal advice on any particular matter.

Contact Details

ADL SOLICITORS
13 ST SWITHINGS LANE
LONDON
EC4N 8AL

0200 888 0250 (+44 200 888 0250)
www.ADLSOLICITORS.COM

nick@adlsolicitors.com

CAAIRO - CENTRE FOR THE ASSESSMENT OF ARTIFICIAL INTELLIGENCE RISKS AND OPPORTUNITIES
14 STAFFORD PLACE, WESTMINSTER
LONDON SW1E
0796 249 8000 (+44 796 249 8000)
www.CAAIRO.COM
www.THETHINKINGMACHINE.UK
nick@thethinkingmachine.uk

Author: DR NICK LOCKETT

Nick Lockett is a barrister and a higher rights solicitor advocate and former investment financier. He has specialized in commercial technology and IT for over 30 years, having been the first barrister in Europe to specialize and write on internet law in 1992 and having been in the AI field for over 10 years. He has been a member of numerous international developments as well as EU projects and practices from offices in the both the City of London and Westminster in London.

Parts of this paper reproduce the Annexes and Glossaries published with the EU AI Act Codes of Practice in their entirety. These are provided and sourced by the European Commission.

Insight Paper 135 An Introduction to the EU's AI Act is also available.

CONTENTS

INTRODUCING THE EU AI ACT

Readers are referred to the Insight Series Paper 125 which introduces the EU AI Act.

The AI Act applies to "AI systems", which the Act defines as:
"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".

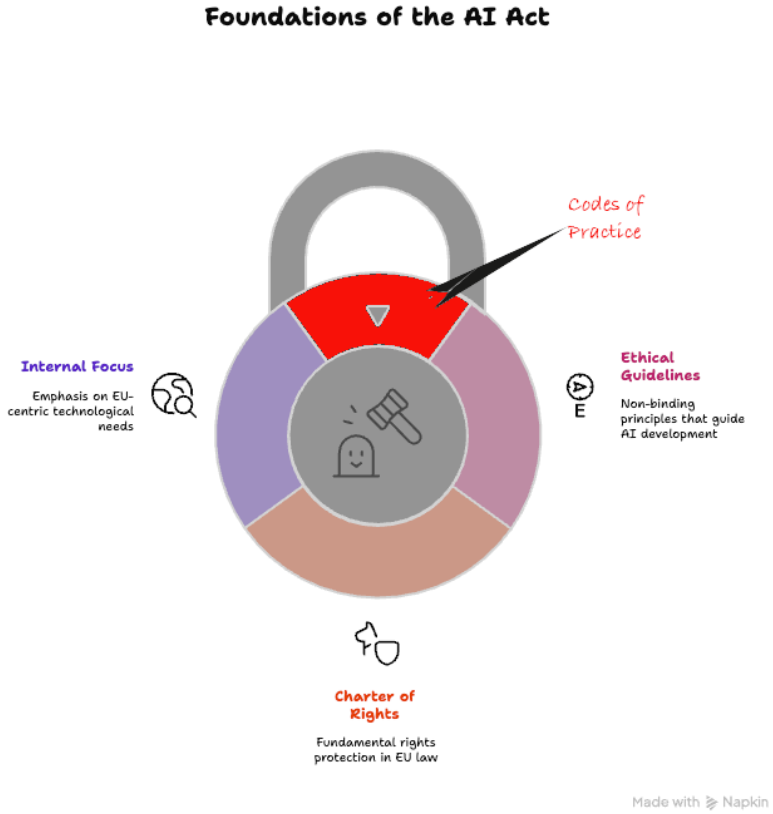
The **EU Artificial Intelligence Act (AI Act)** was officially published in the Official Journal of the EU on July 12, 2024, and entering into force on August 1, 2024, its primary objective is to foster trustworthy AI in Europe by addressing the risks associated with AI systems while simultaneously promoting innovation and ensuring the protection of fundamental rights.

The Act applies to a variety of actors within the AI supply chain, including providers (those who develop AI systems or general-purpose AI models for the EU market), importers (who place AI systems from outside the EU on the market), distributors (who make AI available to others), deployers (entities using a high-risk AI system for professional activities), and manufacturers of products embedding AI systems. While the primary responsibilities often lie with the providers, deployers also bear significant obligations. The Act's extraterritorial reach means that even foreign suppliers must appoint an authorized representative in the Union to ensure compliance if their AI system's output is intended to be used in the EU.

At the core of the EU AI Act is its risk-based regulatory approach, categorizing AI applications into four distinct levels based on the potential harm they could cause to health, safety, or fundamental rights. This tiered system dictates the stringency of the obligations imposed on AI systems.

The AI Act explicitly builds on the Ethical Guidelines on Trustworthy AI, which were published by the European Commission in 2019. While these guidelines remain non-binding, many of their principles have been directly incorporated into the AI Act but like many of the EU laws, they are focussed internally for example are based around a fundamental principle of protecting the Charter of Fundamental Rights of the European Union., rather than starting from an entirely neutral basis to focus on technological needs. As a result, the EU Act contains prohibited AI practices, which cannot be placed on the market or put into service, and using an AI system that employs any of these practices is prohibited.

These guidelines, while non-binding, established a foundational framework for the responsible development and use of AI within the EU. They articulate a vision of "Trustworthy AI" that is rooted in fundamental rights, ethical principles, and technical robustness. This comprehensive analysis delves into the background, core philosophy, and detailed requirements of the EU AI ethical guidelines, providing an in-depth understanding of how the EU seeks to ensure AI is a force for good.


Foundations of the AI Act
Codes of Practice
Internal Focus
Emphasis on EU-centric technological needs
Ethical Guidelines
Non-binding principles that guide AI development
E
Charter of Rights
Fundamental rights protection in EU law
Made with Napkin

I. The Context and Genesis of the EU AI Ethical Guidelines

The EU's foray into AI ethics was driven by a commitment to European values and a recognition of the need for a unified approach to emerging technologies. In 2018, the European Commission launched its AI Strategy, aiming to boost the EU's technological capacity, prepare for socio-economic changes, and establish a robust ethical and legal framework. The AI HLEG, an independent group comprising academics, industry representatives, and civil society experts, was tasked with drafting the Ethics Guidelines, premised on a critical understanding: that Trustworthy AI is not merely a technical concept but a socio-technical one that must be integrated into the entire AI system lifecycle, from design and development to deployment and use.

II. Defining Trustworthy AI: The Three Pillars

The Ethics Guidelines define Trustworthy AI as having three essential components, all of which must be met throughout an AI system's life cycle:

1. **Lawful AI:** AI systems must comply with all applicable laws and regulations, both international, European, and national.
2. **Ethical AI:** AI systems must adhere to fundamental ethical principles and values.
3. **Robust AI:** AI systems must be technically sound and resilient, while also considering their social environment.

The guidelines focus primarily on the second and third components—ethical and robust AI—as the foundation for achieving trustworthy systems because the EU AI Act was due. While lawfulness is a prerequisite, the guidelines recognize that legality alone does not guarantee ethical outcomes or societal well-being.

III. The Ethical Imperatives: Foundational Principles

The guidelines are grounded in four core ethical principles derived from fundamental rights, which serve as the moral compass for the development and use of AI:

A. Respect for Human Autonomy

AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. The guidelines emphasize that AI should act as an enabler, not a replacement for human agency. This principle is crucial in contexts where AI might influence or manipulate human behavior, ensuring that individuals retain control over their choices and are aware when they are interacting with an AI system.

B. Prevention of Harm

This principle mandates that AI systems should not cause or exacerbate harm to human beings, including physical, psychological, or economic harm. It extends to the natural environment and other living beings. Developers are urged to assess the safety of their technologies, minimize unintended consequences, and prevent unacceptable harm. This includes addressing the potential for bias and discrimination that can lead to systemic harm.

C. Fairness

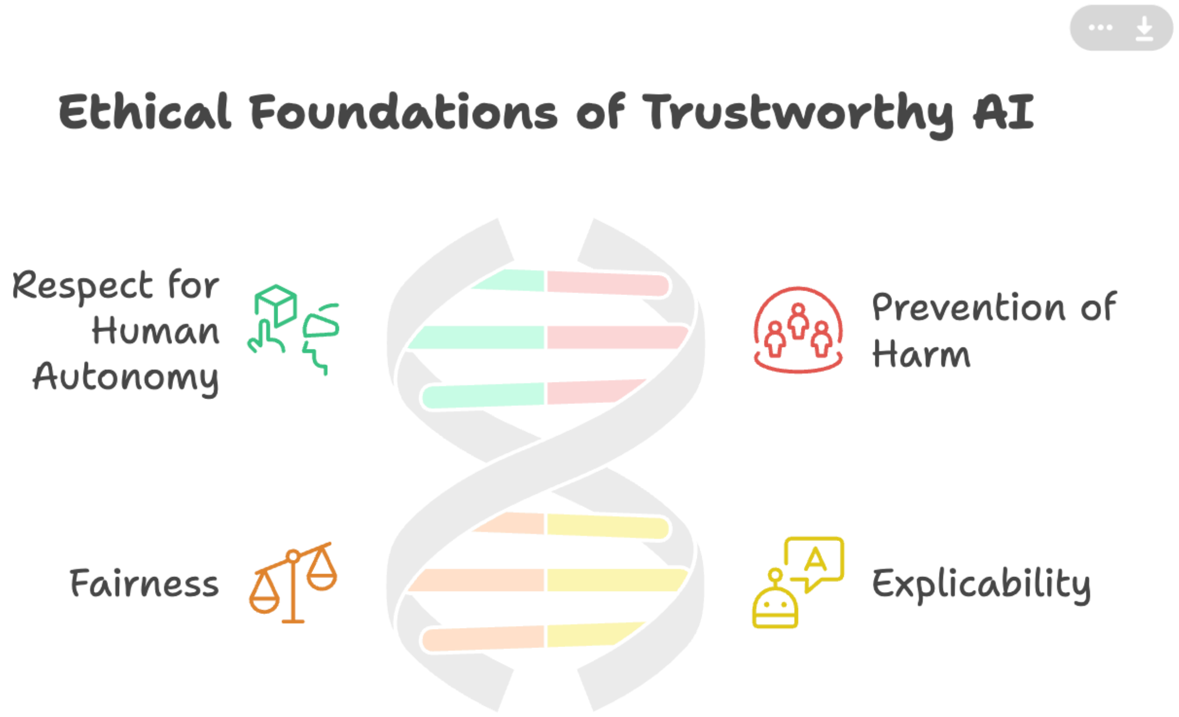
Fairness in AI is a multifaceted concept. It requires ensuring that AI systems avoid unfair bias and discrimination, providing equitable access and treatment for all individuals and groups. The guidelines emphasize that AI should be accessible to all, including marginalized or vulnerable populations, and that the benefits and costs of AI deployment should be fairly distributed.

D. Explicability (Transparency and Explainability)

This principle stresses the importance of transparency in AI systems. Users, developers, and regulators must be able to understand the capabilities, limitations, and decision-making processes of AI. While "explainability" can vary depending on the complexity of the system, the guidelines require a level of understanding that is appropriate for the context and the stakeholders involved. This includes facilitating traceability and auditability, especially in critical applications.

Obligations can be imposed on six categories of economic actors: providers, importers, distributors, product manufacturers, authorised representatives and deployers of GPAI and GPAISR models, all of whom are referred to as “Signatories” in the Codes and are referred to in the commentary as “Providers”.

- Economic operators involved with high-risk AI systems have significant obligations.
- Providers and deployers of certain categories of AI systems are also subject to transparency obligations.
- Providers of general-purpose AI models are subject to obligations.
- The AI Act applies when an AI system or general-purpose AI model is placed on the EU market, put into service in the EU, imported into or distributed in the EU.
- It also applies where an AI system is used by a deployer who has their place of establishment or is in the EU.



Ethical Foundations of Trustworthy AI

IV. The Seven Key Requirements for Trustworthy AI
The guidelines also detail seven key requirements that AI systems should meet, with the aim to provide a practical framework to assess and implement trustworthy AI.

1. Human Agency and Oversight

This requirement reinforces the principle of human autonomy. AI systems should support, rather than diminish, human decision-making and fundamental rights.

- **Human Agency:** Users should be empowered to make informed, autonomous decisions regarding AI systems. This includes having the necessary knowledge and tools to comprehend the system and, where possible, challenge its outcomes.
- **Human Oversight:** Mechanisms must be in place to ensure AI systems do not undermine human autonomy or cause adverse effects. This can be achieved through different models of oversight:
  - **Human-in-the-loop (HITL):** A human directly intervenes in every AI decision.
  - **Human-on-the-loop (HOTL):** A human monitors the AI system's operation and intervenes only when necessary.
  - **Human-in-command (HIC):** A human oversees the overall activity and impact of the AI system and retains the ability to decide when and how to use the system.

2. Technical Robustness and Safety

AI systems must be resilient, reliable, and secure. This requirement aims to minimize and prevent both intentional and unintentional harm.

- **Resilience to Attack and Security:** AI systems must be designed to withstand malicious attacks, including attempts to manipulate data, compromise integrity, or exploit vulnerabilities.
- **Accuracy, Reliability, and Reproducibility:** Systems should consistently produce reliable results across various inputs and situations. Reproducibility is essential for scrutinizing the system and preventing unintended harms.
- **Safety and Fallback Plans:** Developers must implement safeguards and fallback plans to ensure the system's safety in case of errors or failures. This includes minimizing unintended consequences and ensuring the system performs as intended without harming humans or the environment.

3. Privacy and Data Governance

Given that AI systems heavily rely on data, robust privacy protection and ethical data governance are paramount. This requirement aligns closely with the principles of the GDPR.

- **Respect for Privacy:** AI systems must ensure the protection of personal data throughout the entire lifecycle, adhering to principles such as data minimization, pseudonymization, and encryption.
- **Data Quality and Integrity:** Mechanisms for adequate data governance must be established, focusing on the quality, integrity, and ethical sourcing of data used to train and operate AI systems.
- **Legitimized Access to Data:** Clear protocols must be established for accessing and processing data, ensuring that access is controlled and rights-based.

4. Transparency

Transparency ensures that the data, processes, and business models of AI systems are understandable and traceable.

- **Traceability:** It should be possible to trace the AI system's development process and the data used, especially in critical contexts, to identify the cause of errors or unintended outcomes.
- **Explainability:** AI systems and their decisions must be explained in a manner appropriate to the stakeholders concerned. This involves providing clear information about the system's capabilities and limitations.
- **Communication:** Users must be aware that they are interacting with an AI system and understand its functionalities, enabling realistic expectation setting.

5. Diversity, Non-discrimination, and Fairness

Building on the institutional aim of diversity in the EU, this requirement aims to prevent and mitigate unfair bias and ensure that AI systems are inclusive and accessible to all.

- **Avoiding Unfair Bias:** Developers must proactively identify and mitigate biases in data and algorithms that could lead to discrimination or the marginalization of vulnerable groups.
- **Accessibility and Universal Design:** AI systems should be accessible to all users, regardless of disabilities or specific needs, reflecting a commitment to universal design principles.
- **Stakeholder Participation:** Fostering diversity involves engaging relevant stakeholders throughout the entire AI lifecycle to ensure their concerns and perspectives are incorporated.

6. Societal and Environmental Well-being

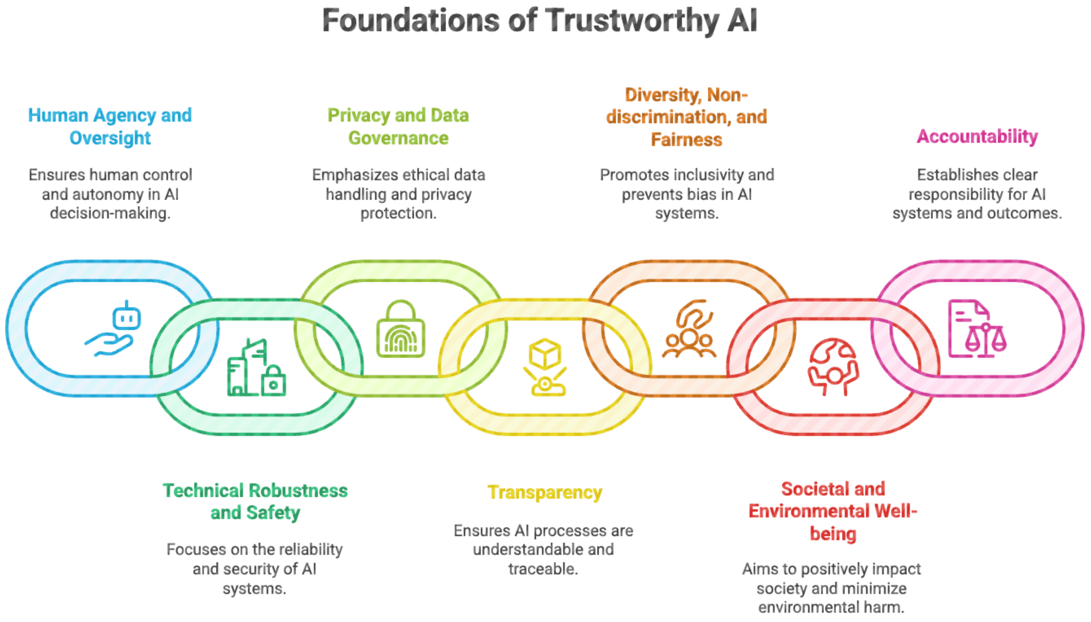
There is an overriding ethos in the EU that AI systems should contribute positively to society and minimize their environmental footprint.

- **Sustainability and Environmental Friendliness:** AI systems should be developed and deployed with consideration for their energy consumption and environmental impact.
- **Social Impact:** The broader societal impact of AI, including effects on democracy, social cohesion, employment, and the rule of law, must be carefully considered and addressed.
- **Benefiting Future Generations:** AI development should be sustainable and ensure that the benefits extend to future generations.

7. Accountability

Accountability ensures that responsibility for AI systems and their outcomes is clearly defined and verifiable.

- **Auditability:** Mechanisms should be in place to enable the assessment of algorithms, data, and design processes, especially in critical applications.
- **Documentation:** Comprehensive documentation of the AI system's development, testing, and deployment is essential for establishing accountability.
- **Redress:** Adequate and accessible mechanisms for redress must be available to individuals who have been harmed by AI systems.



Foundations of Trustworthy AI

Human Agency and Oversight

Ensures human control and autonomy in AI decision-making.

Privacy and Data Governance

Emphasizes ethical data handling and privacy protection.

Diversity, Non-discrimination, and Fairness

Promotes inclusivity and prevents bias in AI systems.

Accountability

Establishes clear responsibility for AI systems and outcomes.

Technical Robustness and Safety

Focuses on the reliability and security of AI systems.

Transparency

Ensures AI processes are understandable and traceable.

Societal and Environmental Well-being

Aims to positively impact society and minimize environmental harm.

Implementation and Assessment: The ALTAI Framework

In an attempt to translate these abstract guidelines into practical application, the Assessment List for Trustworthy AI (ALTAI) was developed.

ALTAI is a self-assessment checklist designed to help developers and deployers of AI implement the seven key requirements in practice and was hoped to provide a structured approach for organizations to evaluate their AI systems against the ethical guidelines similar to ATLAS.

ALTAI stands for the Assessment List for Trustworthy Artificial Intelligence. It is an interactive online prototype developed by the European Commission's High-Level Expert Group on Artificial Intelligence.
Its main purpose is to help assess whether an AI system (during development, deployment, procurement, or use) complies with the seven requirements of Trustworthy AI, as outlined in the Ethics Guidelines for Trustworthy AI.
ALTAI aims to:

- Provide a basis for self-evaluation of Trustworthy AI.
- Help organizations understand the concept and identify potential risks.
- Raise awareness of AI's impact.
- Promote stakeholder involvement.
- Foster responsible and sustainable AI innovation in Europe by integrating ethics into AI development.

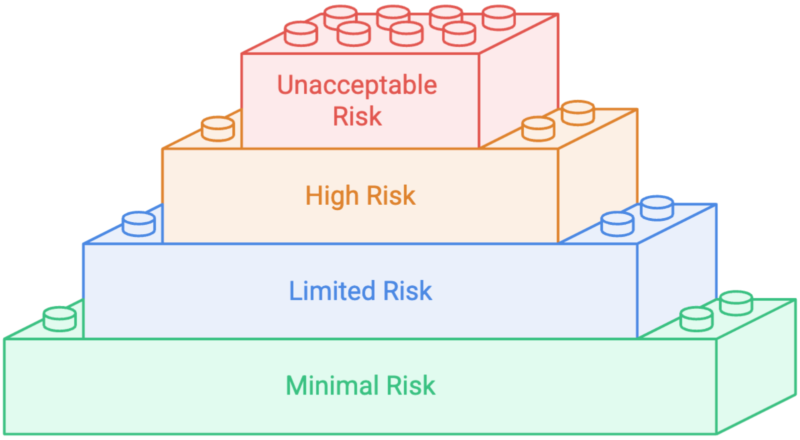
Unfortunately, as The Future Society pointed out, the lobbying of key players in the AI field has resulted in the guidelines being materially weakened despite the hope that this would be a continuous identification, evaluation, and improvement tool that would be effective throughout the AI system’s lifecycle and instead, the need to identify potential risks, document mitigation strategies, and ensure ongoing monitoring has been significantly watered down, as indeed it has within the EU AI Act, although the guidelines emphasize a risk-based approach, where the level of scrutiny and required safeguards are proportionate to the potential risks posed by the AI system, something that was incorporated into the EU AI Act and the Codes of Practice.

Further Reading:
https://altai.insight-centre.org/
https://tinyurl.com/EUEthicsGuidelines
https://corporateeurope.org/sites/default/files/2023-02/The%20Lobbying%20Ghost%20in%20the%20Machine.pdf

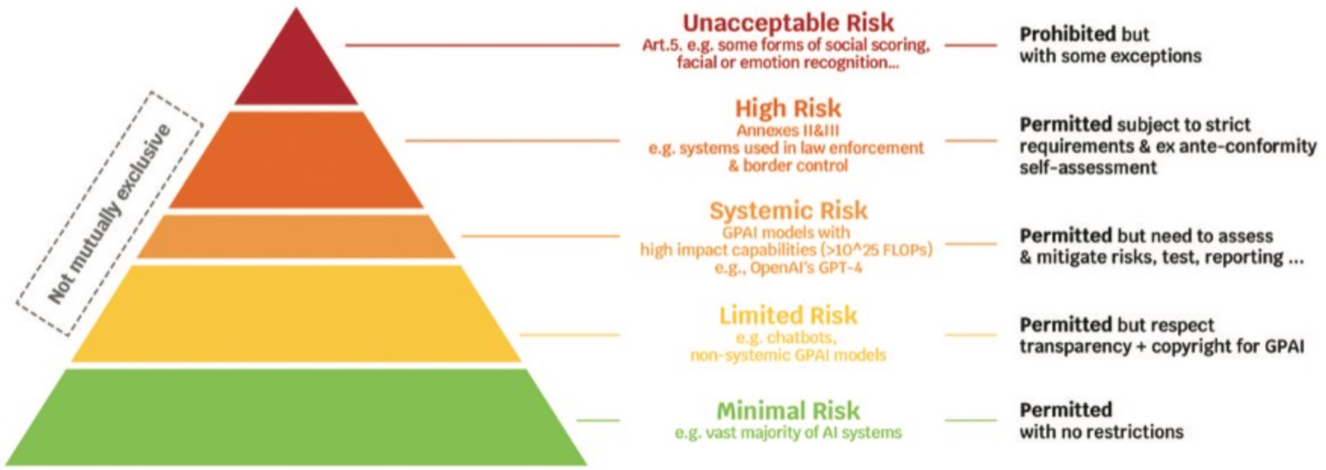
Risk, the Ethical Guidelines and the EU AI Act: A Symbiotic Relationship

The Ethics Guidelines for Trustworthy AI are non-binding recommendations, although they have influenced the regulatory landscape of the EU and served as an ethical bedrock for the **EU AI Act**. The EU AI Act adopts the risk-based approach championed by the guidelines, categorizing AI systems into different risk levels:
Unacceptable Risk.
High Risk.
Limited Risk.
Minimal Risk.

AI Risk Categorization



The EU's push for AI ethics is rooted in a desire to harness the benefits of AI while safeguarding fundamental rights and European values, pushing the EU Agenda in relation to fundamental human rights.


Not mutually exclusive
Unacceptable Risk
Art.5. e.g. some forms of social scoring, facial or emotion recognition...
Prohibited but with some exceptions
High Risk
Annexes II&III
e.g. systems used in law enforcement & border control
Permitted subject to strict requirements & ex ante-conformity self-assessment
Systemic Risk
GPAI models with high impact capabilities (>10^25 FLOPs)
e.g., OpenAI's GPT-4
Permitted but need to assess & mitigate risks, test, reporting ...
Limited Risk
e.g. chatbots, non-systemic GPAI models
Permitted but respect transparency + copyright for GPAI
Minimal Risk
e.g. vast majority of AI systems
Permitted with no restrictions

EU AI ACT – RISK CATEGORIES

Building on the Ethical Guidelines for AI, the EU AI act has adopted the same categories in AI risk development:
1. Prohibited or Unacceptable risk: AI systems that pose such significant risks are unacceptable and therefore prohibited.
2. High risk: High-risk AI systems are subject to stringent regulatory requirements. For example, high-risk AI systems are those that have a significant harmful impact on the health, safety and fundamental rights of persons in the Union.
3. Limited risk: AI systems in this category pose a limited risk, but have specific transparency obligations.
4. Minimal or no risk: AI systems that pose minimal or no risk have no regulatory restrictions under the AI Act.

FURTHER READING

- **The Commission published guidelines on prohibited AI practices on 4 February 2025 available at https://tinyurl.com/EUProhibitedPractices**
- **The AI ACTION SUMMIT INTERNATIONAL AI SAFETY REPORT January 2025** https://tinyurl.com/AIInternationalSafety

THE EU AI ACT

The primary objective of the AI Act is to promote the uptake of human-centric and trustworthy AI within the EU's internal market. It seeks to achieve a delicate balance: encouraging the development and deployment of beneficial AI technologies while mitigating the associated risks to health, safety, and fundamental rights enshrined in the Charter of Fundamental Rights of the European Union.

The AI act applies not only within the EU but potentially beyond the borders of the EU, just as the GDPR did and many of its provisions apply regardless of whether the providers are established or located within the EU or within a third country. The EU AI Act applies to any provider or entity responsible for deploying an AI system if the output produced by the system is intended to be used in the EU, and if it does, then foreign suppliers must appoint an authorised representative in the union to ensure compliance with the acts provisions; however to comply with international legislative norms, the EU AI Act does not apply to public authorities of third countries or international organisations under police and judicial cooperation agreements with the union, no to AI systems placed on the market for military defence or national security purpose.

The EU has long accepted the lack of any ability of regulations to keep up with the pace of change in technology, due to the amount of time it takes to complete regulatory changes. As a result, the EU determine that there would be a combination of regulations together with codes of conduct, that are designed so that by a daring to the code of conduct there is deemed compliance with the EUAI act requirements.

The Act applies broadly to providers, deployers, importers, and distributors of AI systems, regardless of where they are located, provided their AI systems are placed on the EU market or their output is used within the EU. This "extraterritorial effect" ensures that foreign companies operating in the EU market are subject to the same regulatory standards as their European counterparts.

The AI Act defines an "AI system" as a "machine-based system that is 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."

The EU rules for GPAI enter into force in August 2025. In theory, providers may opt out of the Codes for other means to comply with the obligations under the AI Act. However, they would need to provide additional supporting evidence and might be subject to more requests of information, as suggested by the European Commission. In contrast, the Codes of Practice are the most straightforward and transparent way of complying with the AI Act.

II. THE RISK-BASED APPROACH: THE CORE OF THE REGULATION

The AI Act employs a tiered, risk-based approach, which is the cornerstone of the regulation. This system classifies AI systems into four categories based on the level of risk they pose to society and individuals: Unacceptable Risk, High Risk, Limited Risk, and Minimal/No Risk. The obligations imposed on AI providers and deployers are directly proportionate to the level of risk identified.

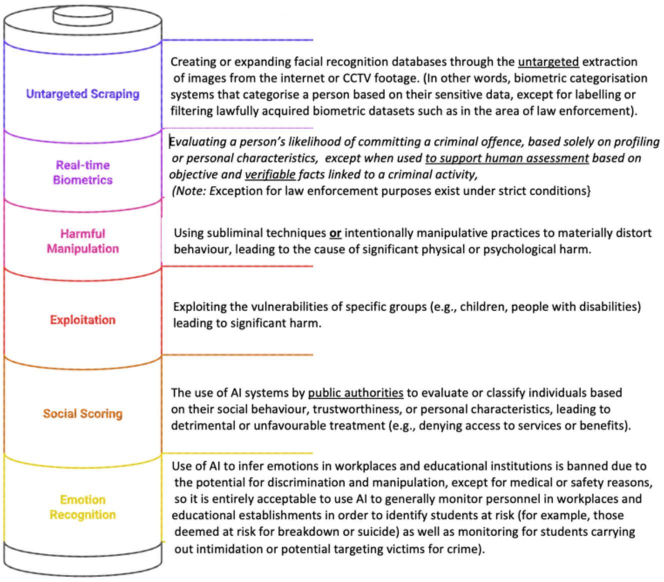
A. Unacceptable Risk: Prohibited Practices

The AI Act takes a decisive stance against AI systems deemed to pose an "unacceptable risk" to fundamental rights and safety. These systems are outright banned within the EU. The prohibition applies to practices that are fundamentally contrary to European values and democratic principles.
Key prohibited practices include:

- **Harmful manipulative techniques:** AI that uses subliminal techniques or intentionally manipulative practices to materially distort behaviour, leading to the cause of significant physical or psychological harm.
- **Exploitation of vulnerabilities:** AI systems that exploit the vulnerabilities of specific groups (e.g., children, people with disabilities) leading to significant harm.
- **Social Scoring:** The use of AI systems by public authorities to evaluate or classify individuals based on their social behaviour, trustworthiness, or personal characteristics, leading to detrimental or unfavourable treatment (e.g., denying access to services or benefits).
- **Real-time remote biometric identification in public spaces:** E*valuating a person's likelihood of committing a criminal offence, based solely on profiling or personal characteristics, except when used to support human assessment based on objective and verifiable facts linked to a criminal activity, so whilst* exceptions for law enforcement purposes exist under strict conditions (e.g., searching for victims of crime, preventing terrorist attacks, preventing crime by identifying prolific offenders so that they can be monitored by AI for suspicious behaviour when they will be alerted to human support personnel), the general use of real-time facial recognition in public spaces is prohibited, although this has been subject to significant criticism.
- **Untargeted scraping of facial images:** Creating or expanding facial recognition databases through the untargeted extraction of images from the internet or CCTV footage. (In other words, biometric categorisation systems that categorise a person based on their sensitive data, except for labelling or filtering lawfully acquired biometric datasets such as images in the area of law enforcement).
- **Emotion recognition in workplaces and educational institutions:** While allowed in limited circumstances, the use of AI to infer emotions in these settings is banned due to the potential for discrimination and manipulation, except for medical or safety reasons, so it is entirely acceptable to use AI to generally monitor personnel in workplaces and educational establishments in order to identify students at risk (for example, those deemed at risk for breakdown or suicide) as well as monitoring for students carrying out intimidation or potential targeting victims for crime).

In some cases, the AI Act contains exceptions that allow these "prohibited" practices to be used in certain situations. A good example is real-time biometric identification, where the Regulation allows its use in exceptional circumstances. The application of these exceptions requires notifications or prior authorisations.

Prohibited Actions



Creating or expanding facial recognition databases through the untargeted extraction of images from the internet or CCTV footage. (In other words, biometric categorisation systems that categorise a person based on their sensitive data, except for labelling or filtering lawfully acquired biometric datasets such as in the area of law enforcement).

Evaluating a person's likelihood of committing a criminal offence, based solely on profiling or personal characteristics, except when used to support human assessment based on objective and verifiable facts linked to a criminal activity,
(Note: Exception for law enforcement purposes exist under strict conditions}

Using subliminal techniques **or** intentionally manipulative practices to materially distort behaviour, leading to the cause of significant physical or psychological harm.

Exploiting the vulnerabilities of specific groups (e.g., children, people with disabilities) leading to significant harm.

The use of AI systems by public authorities to evaluate or classify individuals based on their social behaviour, trustworthiness, or personal characteristics, leading to detrimental or unfavourable treatment (e.g., denying access to services or benefits).

Use of AI to infer emotions in workplaces and educational institutions is banned due to the potential for discrimination and manipulation, except for medical or safety reasons, so it is entirely acceptable to use AI to generally monitor personnel in workplaces and educational establishments in order to identify students at risk (for example, those deemed at risk for breakdown or suicide) as well as monitoring for students carrying out intimidation or potential targeting victims for crime).

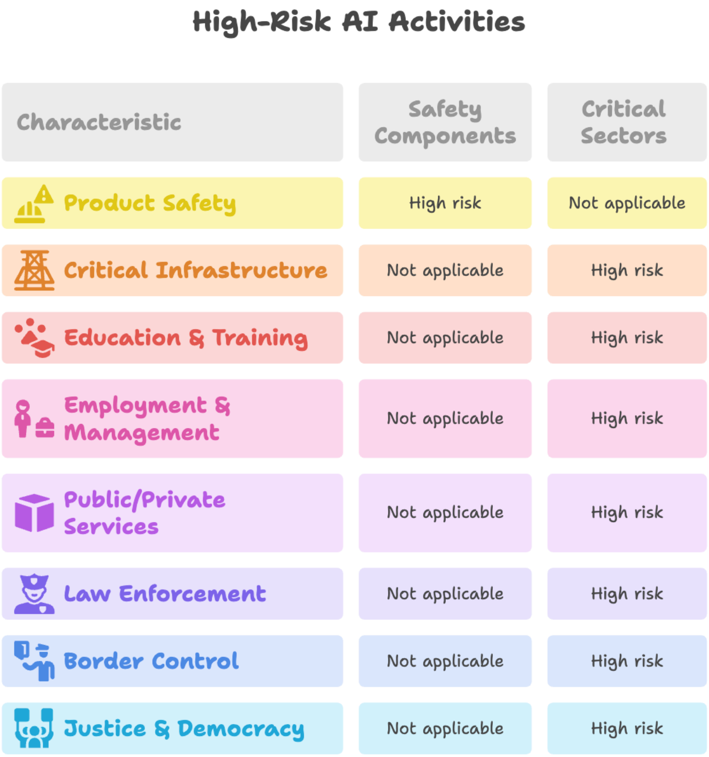
The Commission published guidelines on prohibited AI practices on 4 February 2025 available at https://tinyurl.com/EUProhibitedPractices .
see also https://tinyurl.com/IllusionOfThinking

B. High Risk: Strict Obligations

AI systems classified as "high risk" are permitted but subject to rigorous compliance requirements before they can be placed on the market or put into service. High-risk systems are those that have the potential to pose significant harm to health, safety, or fundamental rights.

The Act identifies high-risk systems through two main pathways:

1. **Safety Components of Products:** AI systems intended to be used as safety components of products already regulated by EU product safety legislation (e.g., medical devices, aviation, automobiles).
2. **Specific Use Cases in Critical Sectors:** AI systems used in specific sectors where the potential impact on fundamental rights is significant. These sectors include:
   - **Critical Infrastructure:** Management and operation of essential services like water, gas, electricity, and transport.
   - **Education and Vocational Training:** Systems determining access to education or assessing student performance.
   - **Employment and Worker Management:** Tools for recruitment, selection, evaluation, or monitoring of workers.
   - **Access to Essential Public and Private Services:** Systems used for credit scoring, evaluating eligibility for public benefits, or dispatching emergency services.
   - **Law Enforcement:** AI used for analyzing evidence, risk assessments, or predictive policing.
   - **Migration, Asylum, and Border Control:** Systems used for verifying visa applications or managing migration flows.
   - **Administration of Justice and Democratic Processes:** AI assisting judicial authorities in preparing court rulings or influencing electoral processes.



High-Risk AI Activities

Characteristic	Safety Components	Critical Sectors
Product Safety	High risk	Not applicable
Critical Infrastructure	Not applicable	High risk
Education & Training	Not applicable	High risk
Employment & Management	Not applicable	High risk
Public/Private Services	Not applicable	High risk
Law Enforcement	Not applicable	High risk
Border Control	Not applicable	High risk
Justice & Democracy	Not applicable	High risk

Obligations for High-Risk AI Systems:
Providers of high-risk AI systems bear the heaviest compliance burden. They must adhere to a comprehensive set of requirements:

1. **Risk Management System:** Establishing and maintaining a continuous risk management system throughout the AI system's lifecycle to identify, analyze, and mitigate potential risks.
2. **Data Governance and Quality:** Implementing rigorous data governance practices to ensure the quality, relevance, and representativeness of training, validation, and testing data. This is crucial for minimizing bias and ensuring accuracy.
3. **Technical Documentation and Record-Keeping:** Maintaining detailed technical documentation and automatically generated logs (log-keeping) to ensure transparency and traceability.
4. **Transparency and Information to Deployers:** Providing clear, comprehensive information to deployers (users) about the AI system's intended purpose, capabilities, limitations, and required human oversight.
5. **Human Oversight:** Designing the system to allow for effective human oversight, ensuring that a human can intervene, interpret the AI's output, and ultimately override automated decisions.
6. **Accuracy, Robustness, and Cybersecurity:** Ensuring a high level of accuracy, technical robustness, and resilience against errors, failures, and cyberattacks.
7. **Conformity Assessment:** High-risk AI systems must undergo a conformity assessment procedure before being placed on the market. Depending on the system, this may involve self-assessment by the provider or a third-party assessment by a notified body.

C. Limited Risk

AI systems in the limited risk category pose lower risks and are primarily subject to transparency obligations. This category includes systems where the potential for manipulation or deception exists, but the risk is not considered high.

The main requirement for limited risk AI systems is that users must be clearly informed when they are interacting with an AI system. Examples include:

- **Chatbots:** Users must be informed that they are communicating with an AI.
- **Emotion recognition systems:** While heavily restricted in certain areas (e.g., workplaces), where used, users must be notified.

- **Deepfakes (AI-generated or manipulated content):** The Act requires that AI-generated synthetic content (images, audio, or video) be clearly labeled as such to prevent manipulation.

D. Minimal Risk

The vast majority of AI systems, such as spam filters, AI-enabled search engines, and video games, fall into this category. Although the EU AI Act uses the term no-risk, this is considered a misnomer as all systems carry risk and the correct assessment is minimal risk. The AI Act imposes no mandatory requirements on these systems. However, the Act encourages adherence to voluntary codes of conduct to promote ethical principles and best practices.

Following its commitment at the Bletchley Summit, the UK Government has published the AI Action Summit's International AI Safety Report (https://tinyurl.com/AIINTERNATIONALSAFETY)

REGULATION OF GENERAL PURPOSE AI (GPAI) AND FOUNDATION MODELS

A crucial addition during the legislative process was the regulation of General Purpose AI (GPAI), also known as foundation models (e.g., large language models like GPT-4 or Claude). These models, capable of performing a wide range of tasks and adaptable to various applications, present unique challenges.

The AI Act introduces specific requirements for GPAI models, differentiating between standard GPAI and GPAI with "systemic risks."

OBLIGATIONS FOR GPAI PROVIDERS:

- **Documentation:** Providers must maintain comprehensive technical documentation, including details about training and testing.
- **Transparency:** They must provide information and documentation to downstream providers (those integrating the GPAI model into their own AI systems) to help them understand the model's capabilities and limitations.
- **Copyright Compliance:** Providers must establish policies to comply with EU copyright law, including providing a sufficiently detailed summary of the content used for training the model.

GPAI MODELS WITH SYSTEMIC RISK:

GPAI models are classified as having "systemic risk" if they possess significant computational power (exceeding a defined threshold, currently 1025 FLOPs) or if the European Commission designates them as such due to their widespread impact.

Providers of systemic GPAI models face additional, stringent obligations, including:

- **Model Evaluation:** Conducting rigorous evaluations to assess and mitigate systemic risks.
- **Incident Reporting:** Tracking, documenting, and reporting serious incidents related to the model.
- **Cybersecurity:** Ensuring a high level of cybersecurity protection for the model and its infrastructure.

GOVERNANCE, ENFORCEMENT, AND PENALTIES

The AI Act establishes a robust governance structure for oversight and enforcement.

Governance Structure:

- **European AI Office:** Established within the European Commission, the AI Office is the central body responsible for coordinating enforcement, monitoring the implementation of the Act, and overseeing GPAI models.
- **AI Board:** Comprised of representatives from Member States, the AI Board advises the Commission and helps ensure harmonized application of the Act across the EU.
- **National Competent Authorities:** Member States are responsible for designating market surveillance authorities to enforce the Act at the national level.

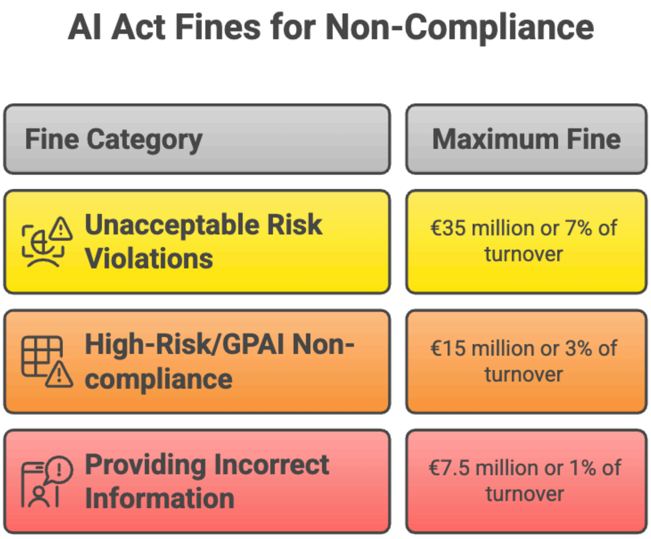
Penalties:

The AI Act introduces substantial fines for non-compliance, comparable to those under the GDPR, reinforcing the seriousness of the regulation.

- **Unacceptable Risk Violations:** Fines can reach up to €35 million or 7% of the company's total worldwide annual turnover, whichever is higher.

- **Non-compliance with High-Risk Requirements or GPAI Obligations:** Fines can reach up to €15 million or 3% of the total worldwide annual turnover.

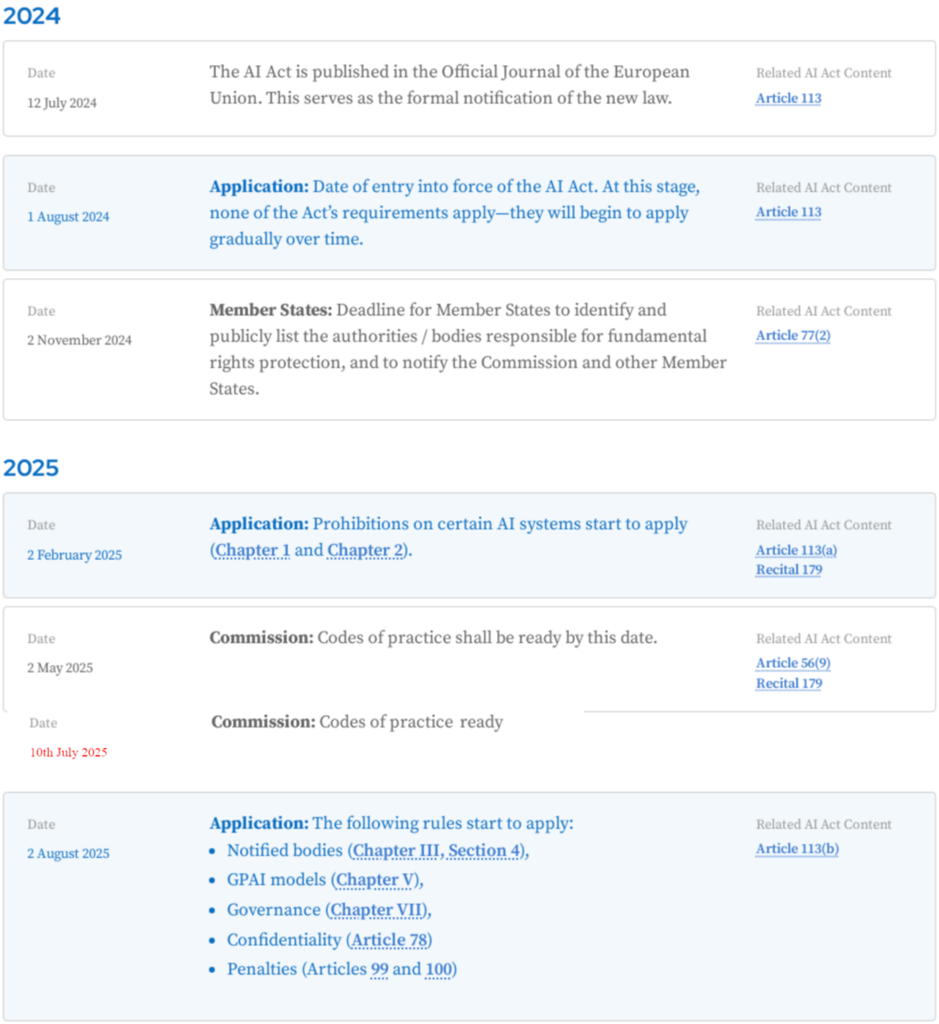
- **Providing Incorrect Information:** Fines can reach up to €7.5 million or 1% of the total worldwide annual turnover.



AI Act Fines for Non-Compliance

Fine Category	Maximum Fine
Unacceptable Risk Violations	€35 million or 7% of turnover
High-Risk/GPAI Non-compliance	€15 million or 3% of turnover
Providing Incorrect Information	€7.5 million or 1% of turnover

TIMETABLES



2024

Date
12 July 2024

The AI Act is published in the Official Journal of the European Union. This serves as the formal notification of the new law.

Related AI Act Content
Article 113

Date
1 August 2024

Application: Date of entry into force of the AI Act. At this stage, none of the Act's requirements apply—they will begin to apply gradually over time.

Related AI Act Content
Article 113

Date
2 November 2024

Member States: Deadline for Member States to identify and publicly list the authorities / bodies responsible for fundamental rights protection, and to notify the Commission and other Member States.

Related AI Act Content
Article 77(2)

2025

Date
2 February 2025

Application: Prohibitions on certain AI systems start to apply (Chapter 1 and Chapter 2).

Related AI Act Content
Article 113(a)
Recital 179

Date
2 May 2025

Commission: Codes of practice shall be ready by this date.

Related AI Act Content
Article 56(9)
Recital 179

Date
10th July 2025

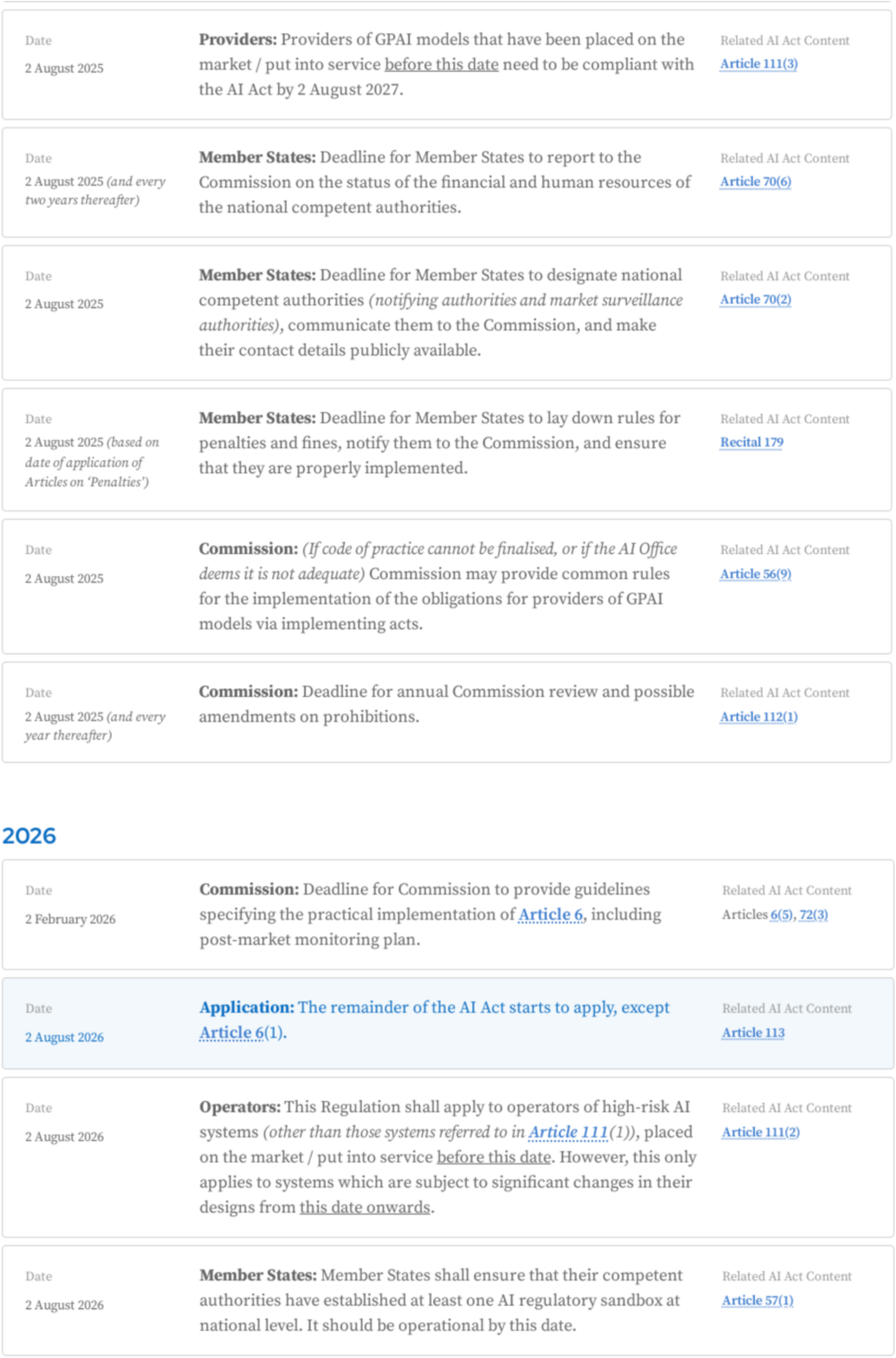
Commission: Codes of practice ready

Date
2 August 2025

Application: The following rules start to apply:

- Notified bodies (Chapter III, Section 4),
- GPAI models (Chapter V),
- Governance (Chapter VII),
- Confidentiality (Article 78)
- Penalties (Articles 99 and 100)

Related AI Act Content
Article 113(b)



Date
2 August 2025

Providers: Providers of GPAI models that have been placed on the market / put into service before this date need to be compliant with the AI Act by 2 August 2027.

Related AI Act Content
Article 111(3)

Date
2 August 2025 *(and every two years thereafter)*

Member States: Deadline for Member States to report to the Commission on the status of the financial and human resources of the national competent authorities.

Related AI Act Content
Article 70(6)

Date
2 August 2025

Member States: Deadline for Member States to designate national competent authorities *(notifying authorities and market surveillance authorities)*, communicate them to the Commission, and make their contact details publicly available.

Related AI Act Content
Article 70(2)

Date
2 August 2025 *(based on date of application of Articles on 'Penalties')*

Member States: Deadline for Member States to lay down rules for penalties and fines, notify them to the Commission, and ensure that they are properly implemented.

Related AI Act Content
Recital 179

Date
2 August 2025

Commission: *(If code of practice cannot be finalised, or if the AI Office deems it is not adequate)* Commission may provide common rules for the implementation of the obligations for providers of GPAI models via implementing acts.

Related AI Act Content
Article 56(9)

Date
2 August 2025 *(and every year thereafter)*

Commission: Deadline for annual Commission review and possible amendments on prohibitions.

Related AI Act Content
Article 112(1)

2026

Date
2 February 2026

Commission: Deadline for Commission to provide guidelines specifying the practical implementation of Article 6, including post-market monitoring plan.

Related AI Act Content
Articles 6(5), 72(3)

Date
2 August 2026

Application: The remainder of the AI Act starts to apply, except Article 6(1).

Related AI Act Content
Article 113

Date
2 August 2026

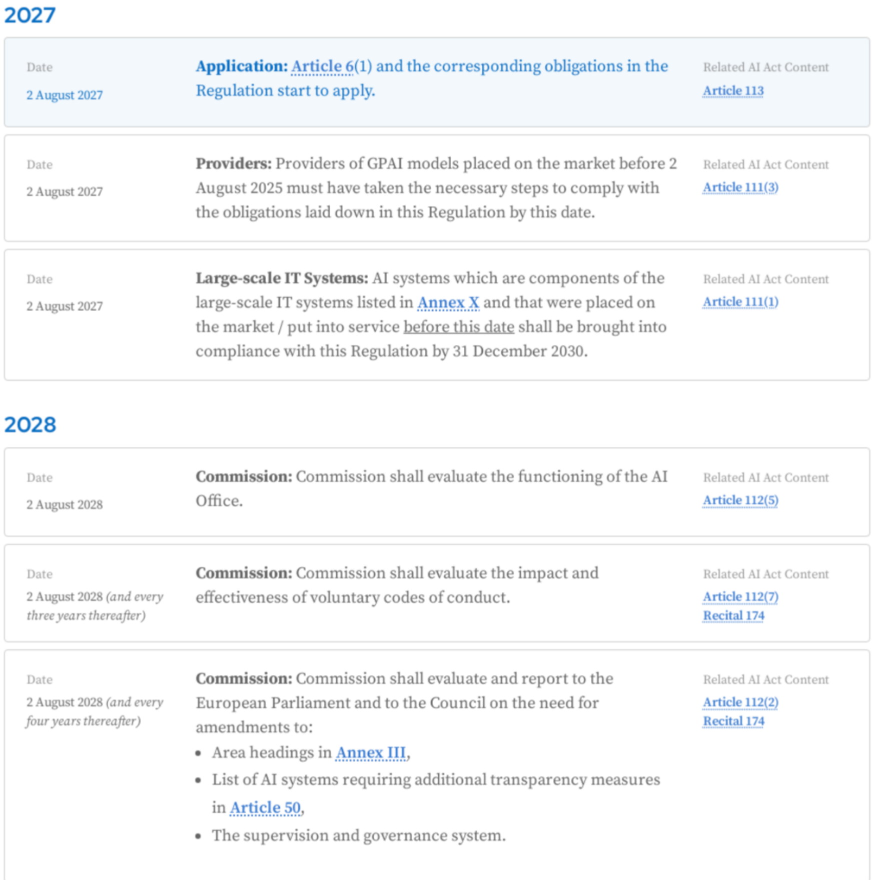
Operators: This Regulation shall apply to operators of high-risk AI systems *(other than those systems referred to in Article 111(1))*, placed on the market / put into service before this date. However, this only applies to systems which are subject to significant changes in their designs from this date onwards.

Related AI Act Content
Article 111(2)

Date
2 August 2026

Member States: Member States shall ensure that their competent authorities have established at least one AI regulatory sandbox at national level. It should be operational by this date.

Related AI Act Content
Article 57(1)



2027

Date
2 August 2027

Application: Article 6(1) and the corresponding obligations in the Regulation start to apply.

Related AI Act Content
Article 113

Date
2 August 2027

Providers: Providers of GPAI models placed on the market before 2 August 2025 must have taken the necessary steps to comply with the obligations laid down in this Regulation by this date.

Related AI Act Content
Article 111(3)

Date
2 August 2027

Large-scale IT Systems: AI systems which are components of the large-scale IT systems listed in Annex X and that were placed on the market / put into service before this date shall be brought into compliance with this Regulation by 31 December 2030.

Related AI Act Content
Article 111(1)

2028

Date
2 August 2028

Commission: Commission shall evaluate the functioning of the AI Office.

Related AI Act Content
Article 112(5)

Date
2 August 2028 *(and every three years thereafter)*

Commission: Commission shall evaluate the impact and effectiveness of voluntary codes of conduct.

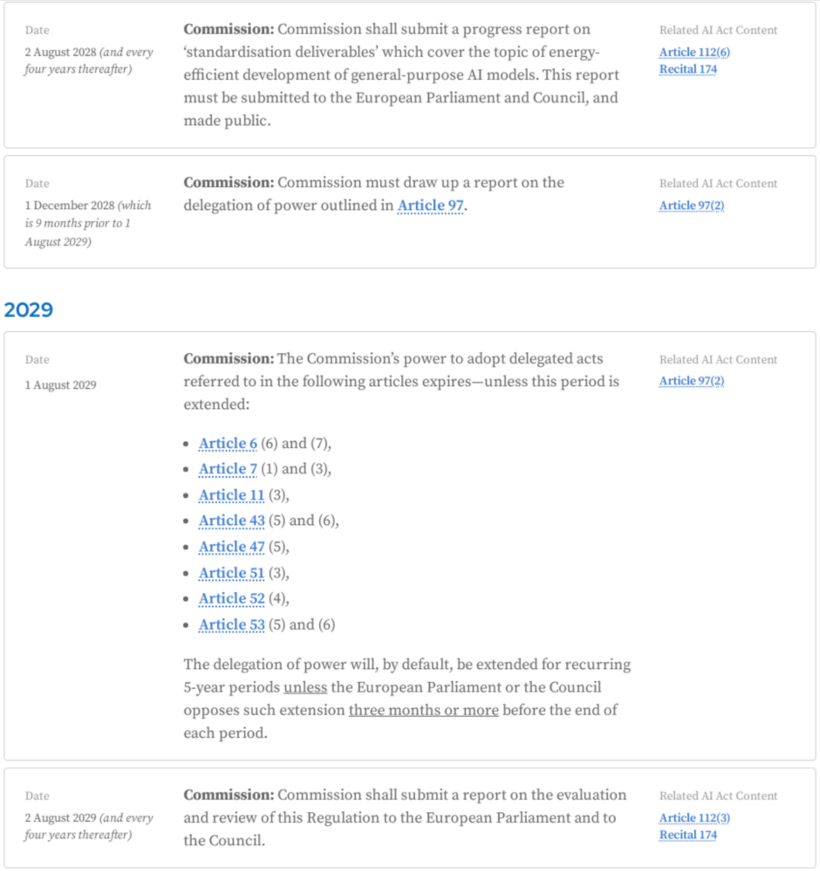
Related AI Act Content
Article 112(7)
Recital 174

Date
2 August 2028 *(and every four years thereafter)*

Commission: Commission shall evaluate and report to the European Parliament and to the Council on the need for amendments to:

- Area headings in Annex III,
- List of AI systems requiring additional transparency measures in Article 50,
- The supervision and governance system.

Related AI Act Content
Article 112(2)
Recital 174



Date
2 August 2028 *(and every four years thereafter)*

Commission: Commission shall submit a progress report on 'standardisation deliverables' which cover the topic of energy-efficient development of general-purpose AI models. This report must be submitted to the European Parliament and Council, and made public.

Related AI Act Content
Article 112(6)
Recital 174

Date
1 December 2028 *(which is 9 months prior to 1 August 2029)*

Commission: Commission must draw up a report on the delegation of power outlined in Article 97.

Related AI Act Content
Article 97(2)

2029

Date
1 August 2029

Commission: The Commission's power to adopt delegated acts referred to in the following articles expires—unless this period is extended:

- Article 6 (6) and (7),
- Article 7 (1) and (3),
- Article 11 (3),
- Article 43 (5) and (6),
- Article 47 (5),
- Article 51 (3),
- Article 52 (4),
- Article 53 (5) and (6)

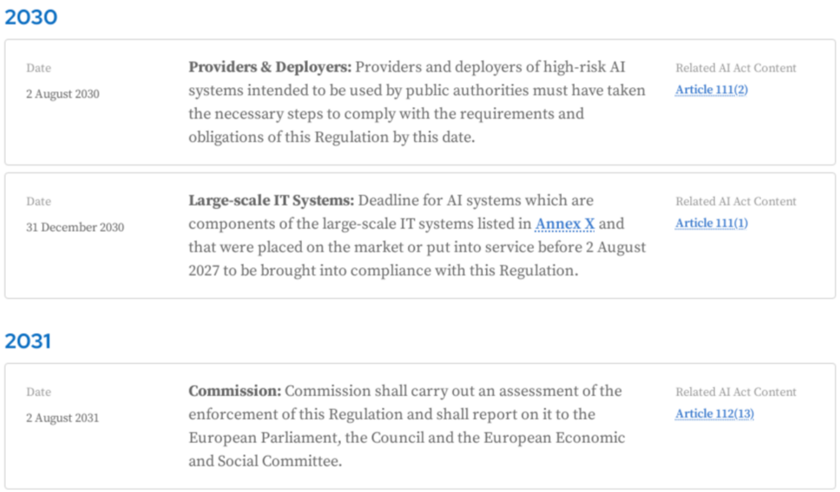
The delegation of power will, by default, be extended for recurring 5-year periods unless the European Parliament or the Council opposes such extension three months or more before the end of each period.

Related AI Act Content
Article 97(2)

Date
2 August 2029 *(and every four years thereafter)*

Commission: Commission shall submit a report on the evaluation and review of this Regulation to the European Parliament and to the Council.

Related AI Act Content
Article 112(3)
Recital 174



2030

Date		Related AI Act Content
2 August 2030	**Providers & Deployers:** Providers and deployers of high-risk AI systems intended to be used by public authorities must have taken the necessary steps to comply with the requirements and obligations of this Regulation by this date.	Article 111(2)

Date		Related AI Act Content
31 December 2030	**Large-scale IT Systems:** Deadline for AI systems which are components of the large-scale IT systems listed in Annex X and that were placed on the market or put into service before 2 August 2027 to be brought into compliance with this Regulation.	Article 111(1)

2031

Date		Related AI Act Content
2 August 2031	**Commission:** Commission shall carry out an assessment of the enforcement of this Regulation and shall report on it to the European Parliament, the Council and the European Economic and Social Committee.	Article 112(13)

NB The timetable may be updated by the European Commission or the EU AI Office

SAFETY COMPONENTS

There are special provisions where GPAIs and GPAISRs are incorporated into safety components or safety products. This covers AI systems intended to be used as a product or a safety component of a product which is covered by EU harmonisation legislation, such as civil aviation, vehicle security, marine equipment, radio equipment, toys, lifts, pressure equipment, medical devices, personal protective equipment[1] and also covers remote biometric identification systems, and AI systems used as a safety component in critical infrastructure.

Safety Components and AI Embedded in Safety Products: A Deep Dive
The EU AI Act's focus on safety and security becomes particularly stringent when AI is integrated into **safety components** or **embedded in safety products**. These are typically categorized as high-risk AI systems due to their direct potential to cause harm to individuals or property if they fail.

"Safety Components" and "AI Embedded in Safety Products"
The Act defines high-risk AI systems in two main categories:

1. **AI systems intended to be used as a safety component of a product, or which are themselves products covered by existing EU harmonization legislation.**
   This includes a wide range of products where AI's failure could lead to significant harm. Examples include:
   - **Medical Devices:**
     AI systems used for diagnosis (e.g., detecting tumors in scans), treatment planning, or monitoring vital signs.
   - **Machinery and Robotics:**
     AI systems controlling industrial robots, automated production lines, or safety functions in heavy machinery.
   - **Automotive Sector:**
     AI systems for autonomous driving (e.g., perception, decision-making, control), advanced driver-assistance systems (ADAS), or safety features like automatic emergency braking.
   - **Aviation:**
     AI used in air traffic control, flight management systems, or drone operations.
   - **Toys and Consumer Products:**
     AI in products where malfunction could pose a physical safety risk.

2. **AI systems falling into specific listed critical areas**
   (e.g., critical infrastructure, education, employment, law enforcement, migration, justice).
   While these also have safety implications, the user's question specifically targets the first category related to physical products and components.

How Safety and Security Requirements Apply to These Products

For AI embedded in safety products or acting as safety components, the general requirements outlined in Section III are amplified and interpreted with the highest degree of rigor:

- **Robust Risk Management (Article 9):**
  The risk management system must be exceptionally thorough, identifying not only foreseeable risks but also potential emergent risks inherent in complex AI behaviour. This includes risks from data shifts, adversarial attacks, and unexpected environmental conditions. Mitigation strategies must prioritize fail-safe mechanisms and graceful degradation.

[1] listed in Annex I to the AI Act

- **Impeccable Data Governance and Quality (Article 10):**
  Given that even subtle data flaws can lead to catastrophic failures in safety-critical applications, the quality of training, validation, and testing data is paramount. Data must be highly representative of real-world operational environments, meticulously cleaned, and continuously monitored for drift. The Act emphasizes that data used for high-risk AI systems must be "appropriate" and free from errors that could lead to discrimination or unsafe outcomes.

- **Uncompromising Accuracy, Robustness, and Cybersecurity (Article 15):**
  - **Accuracy:**
    AI in safety products must demonstrate extremely high levels of accuracy, often requiring performance metrics that exceed human capabilities in specific tasks.
  - **Robustness:**
    Systems must be highly resilient to adversarial attacks (e.g., subtle changes to sensor inputs that could trick an autonomous vehicle), environmental noise, and system failures. This might involve redundant systems, error detection and correction mechanisms, and rigorous testing under stress conditions.
  - **Cybersecurity:**
    Protection against cyber threats is critical. A compromised AI in a medical device or an autonomous vehicle could have life-threatening consequences. Cybersecurity measures must be state-of-the-art, including secure coding practices, vulnerability management, and protection against data manipulation that could alter AI behaviour.

- **Effective Human Oversight (Article 14):**
  For safety components, human oversight is not merely about understanding but about intervention and ultimate control. The design must enable humans to:
  - **Override AI decisions:**
    Allow for human intervention to stop or correct the AI's actions if it behaves unsafely.
  - **Monitor effectively:**
    Provide clear, interpretable information about the AI's state, performance, and confidence levels.
  - **Understand limitations:**
    Ensure operators are fully aware of the AI's operational design domain (ODD) and its limitations.

- **Rigorous Conformity Assessment (Article 43):**
  These systems undergo the most stringent conformity assessment procedures, often involving third-party audits and certification. This ensures independent verification that the AI system meets all safety and security requirements before it can be placed on the market or put into service. The CE marking indicates compliance.

- **Robust Post-market Monitoring (Article 61):**
  Continuous monitoring is vital. Any unforeseen safety incidents, performance degradation, or new vulnerabilities must be promptly identified, investigated, and addressed. This includes mechanisms for reporting serious incidents to market surveillance authorities.

The Interplay between the Code of Practice and AI in Safety Products

While the Code of Practice (CoP) primarily targets providers of General Purpose AI (GPAI) models, its influence extends significantly to AI embedded in safety products. This is because many high-risk AI systems, including those in safety-critical applications, are increasingly built upon or utilize GPAI models as foundational components.

- **Upstream Influence:** If a GPAI model (e.g., a large vision model) is used as a component in an autonomous driving system, the safety and security practices of the GPAI provider directly impact the safety and security of the downstream high-risk product.

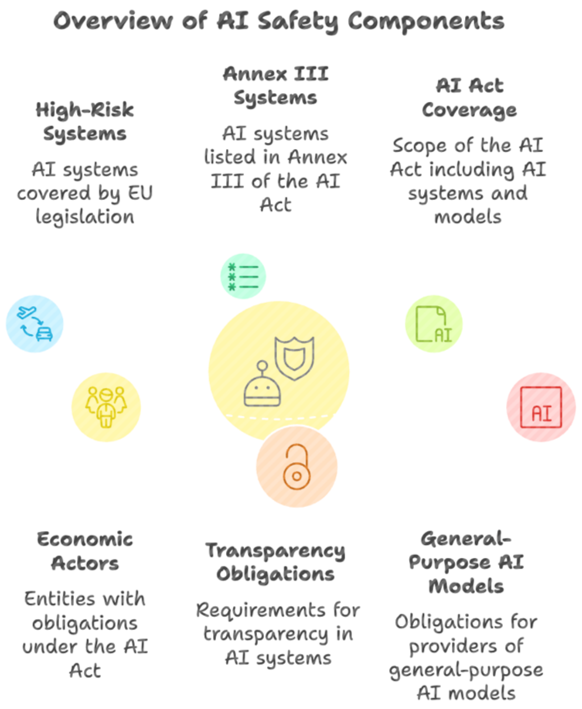
- **Best Practices for Foundational Models:** The CoP, by encouraging GPAI providers to adopt best practices for data governance, robustness, and cybersecurity, indirectly contributes to the safety of high-risk AI systems that build upon these foundational models. For instance, if a GPAI provider ensures their model is trained on diverse, high-quality data and is robust against adversarial attacks (as per Code of Practice guidelines), this reduces the burden and risk for the developer integrating that GPAI model into a safety-critical application.

- **Transparency in the Supply Chain:** The Codes of Practice's emphasis on transparency for GPAI models can provide crucial information to developers of high-risk AI systems. Knowing the training data, known limitations, and evaluation methodologies of the GPAI model allows the high-risk AI provider to conduct their own risk assessments more effectively and tailor their specific safety measures.
- **Mitigation of Systemic Risks:** GPAI models, especially those with systemic risk, could propagate safety or security vulnerabilities across numerous downstream applications. The Code of Practice aims to mitigate these systemic risks at the source, preventing widespread issues in safety-critical products that rely on them.

However, it's crucial to note that adherence to the Code of Practice by a GPAI provider does *not* absolve the provider of a high-risk AI system (or the manufacturer of a safety product embedding AI) from their direct and stringent obligations under the AI Act.

The Code of Practice provides guidance for the foundational layer, but the ultimate responsibility for the safety and security of the final high-risk AI system or product lies with its provider/manufacturer, who must ensure full compliance with all relevant articles of the Act and existing sectoral legislation.

The application of safety and security requirements to AI embedded in safety products presents significant challenges. The dynamic nature of AI, particularly its ability to learn and adapt, makes static safety certification difficult. Ensuring continuous compliance, managing model drift, and defending against sophisticated cyber threats in real-time are ongoing challenges. The sheer complexity of these systems also makes comprehensive testing and validation a monumental task.


Overview of AI Safety Components
High-Risk Systems
AI systems covered by EU legislation
Annex III Systems
AI systems listed in Annex III of the AI Act
AI Act Coverage
Scope of the AI Act including AI systems and models
AI
AI
Economic Actors
Entities with obligations under the AI Act
Transparency Obligations
Requirements for transparency in AI systems
General-Purpose AI Models
Obligations for providers of general-purpose AI models

Bright Spots in AI Legal Governance:

1. **Emergence of Comprehensive Frameworks:** We're seeing a rapid development of dedicated AI risk management frameworks and regulations globally.

    - **EU AI Act:** This is a landmark piece of legislation that takes a risk-based approach, categorizing AI systems into different risk levels (unacceptable, high, limited, minimal) with corresponding obligations. It aims to foster safe and trustworthy AI while ensuring respect for fundamental rights. Its extraterritorial scope means it will influence AI development and deployment far beyond the EU.

    - **NIST AI Risk Management Framework (AI RMF):** In the US, the National Institute of Standards and Technology has published a framework that provides a structured approach to managing AI risks, promoting trustworthy and responsible AI practices. It outlines four core functions: Govern, Map, Measure, and Manage.

    - **ISO/IEC Standards:** International standards like ISO/IEC 27001:2022 and ISO/IEC 23894:2023 offer frameworks for protecting personal data in AI systems and mitigating threats like bias and adversarial attacks. ISO 42001 is also emerging as a key standard for AI management systems.

    - **Pro-Innovation Approach:** Aiming to leverage existing regulators and legal structures while guided by five core principles: safety, security and robustness; appropriate transparency and explainability; fairness; accountability and governance; and contestability and redress, coupled with Flexible Codes of Operation and Conduct that allow fast response to unforeseen circumstances.

2. **Abolition of Downstream Liability Exclusion Clauses**
The Courts walking around the general liability exclusion clauses so that developers can no longer rely upon a general exclusion of downstream liability.

3. **Focus on Key Principles:** Across different jurisdictions, there's a strong consensus on core principles that underpin responsible AI governance:

    - **Transparency and Explainability:** The push for understanding how AI systems work, make decisions, and when they are being used. This builds trust and allows for accountability.
    - **Fairness and Bias Mitigation:** Recognizing and actively working to reduce biases that can be inherited from training data, leading to discriminatory outcomes. This often involves diverse data collection and algorithmic fairness techniques.
    - **Accountability and Governance:** Clearly defining roles, responsibilities, and oversight mechanisms for AI systems throughout their lifecycle. This includes assigning accountability for potential harms.
    - **Safety, Security, and Robustness:** Ensuring AI systems function reliably, securely, and are resilient to attacks or unforeseen issues.
    - **Privacy and Data Protection:** Strict adherence to data protection laws (like GDPR) when AI systems process personal data, including consent, data minimization, and protection against breaches.
    - **Human Oversight and Contestability:** Ensuring that humans remain in control of critical decisions made by AI and providing mechanisms for individuals to challenge or seek redress for harmful AI-driven outcomes.
    -

4. **Cross-Sectoral and International Collaboration:**
    - Many existing laws and regulations (e.g., data protection, consumer protection, anti-discrimination) are being applied or reinterpreted to cover AI.

- There's increasing international dialogue and collaboration (e.g., G7 AI principles, Bletchley Declaration) to develop interoperable approaches and avoid regulatory fragmentation.

5. **Industry-Led Initiatives and Best Practices:** Many companies developing and deploying AI are proactively implementing internal governance frameworks, ethical guidelines, and risk management strategies to build trustworthy AI systems. This includes creating dedicated AI ethics committees and chief AI officer roles. These have not gone far enough and there are inadequate rewards for whistleblowing – just as we spend 5% on NATO spending, companies should be spending at least 5% on Ethics and a further 5% on Safety, but persuading the VC and investors of this will be difficult until the first AI companies is destroyed entirely by class litigation…which will come!

Ways to Govern and Manage AI Risks from a Legal Perspective:

1. **Risk-Based Regulation:** This is a predominant approach, exemplified by the EU AI Act. It means that the level of regulatory scrutiny is proportional to the potential risk an AI system poses. High-risk systems (e.g., in critical infrastructure, healthcare, law enforcement) face more stringent requirements, while lower-risk systems have lighter obligations.
2. **Establishing Clear Definitions and Scopes:** Regulators are working to define what constitutes an "AI system" and to clarify the scope of different regulations to avoid ambiguity and ensure consistent application.
3. **Mandating Risk Assessments and Audits:** Legal frameworks are increasingly requiring organizations to conduct comprehensive risk assessments throughout the AI lifecycle (development, deployment, operation). This includes identifying potential vulnerabilities, biases, and security risks, and implementing mitigation strategies. Regular audits can ensure ongoing compliance.
4. **Transparency and Documentation Requirements:** Laws are pushing for greater transparency about how AI systems are designed, trained, and used. This can involve requirements for detailed documentation of AI models, data sources, and decision-making processes, as well as clear communication to users when they are interacting with AI.
5. **Data Governance and Quality Standards:** Recognizing that AI is only as good as its data, legal frameworks emphasize the importance of high-quality, unbiased, and securely managed data for training and operating AI systems. This includes provisions for data privacy, security, and integrity.
6. **Liability Frameworks:** Work is underway to clarify liability in cases where AI systems cause harm. This is a complex area, but the aim is to ensure that there are clear routes for redress and that accountability can be assigned, whether to developers, deployers, or others in the AI value chain.
7. **Human-in-the-Loop Mechanisms:** For critical AI applications, legal guidance often promotes "human-in-the-loop" oversight, where human judgment and intervention are required at key decision points, especially when there are significant impacts on individuals.
8. **Ethical Guidelines and Principles as Legal Precursors:** Ethical AI principles are often being codified into law, ensuring that ethical considerations are not merely aspirational but become legally enforceable obligations.
9. **Shared Risk Databases such as ATLAS MATRIX**
10. **Thundernerds –** An AI-led "International Risk-U" real-time analysis of the various ethical and risk matrices available worldwide with the aim of providing a comprehensive Universal checklist of what might be the best-practice from the various published risk matrices, project being undertaken by Asimov-Risk and Caairo.

THE CODES OF PRACTICE

NOTE: The Codes were updated on 10th July and are available at https://tinyurl.com/EUAICodesofPractice

The AI Codes of Practice & their criticisms

Although EU AI Act itself is a regulation, it mandates the drawing up of Codes of Practice, particularly for General-Purpose AI (GPAI) models, to help providers comply with their obligations under the Act.

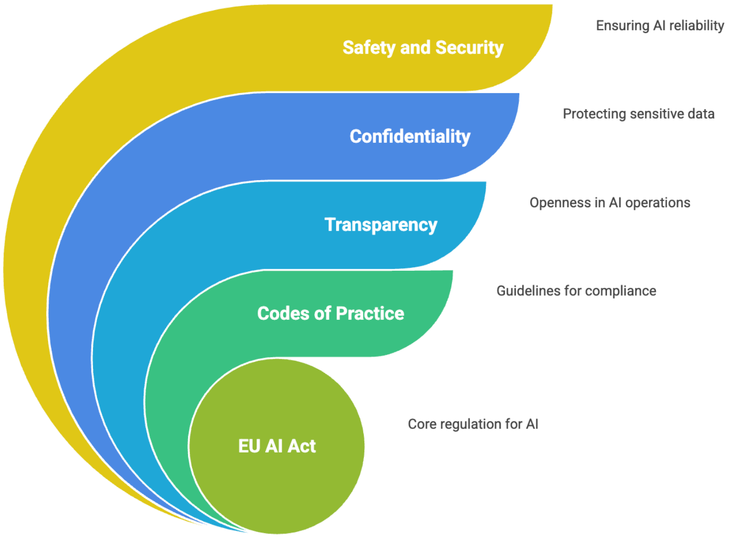
This reflects the fact that the European Commission Expert Groups noted that regulation cannot move fast enough to keep up with fast moving areas such as technology and biotechnology.

These codes are not legally binding in the same way as the Act itself, but adherence to them can provide a "presumption of conformity" with the Act's requirements, particularly when applied alongside protections such as ATLAS Matrix.

There are three Codes:
1. Transparency
2. Confidentiality
3. Safety and Security

To date, the Codes have received criticism from all sides, with model providers saying they will not sign up to the Code as they are unhappy with impractical provisions which go beyond the Act itself, and as they consider is imposes a disproportionate burden on AI providers whilst those on the political left supporting Fundamental Rights consider that the Code does not go far enough and needs to place more burdens on AI Providers.

Although changes have been made to this final version in an attempt to address some of these concerns, whether this is enough to win around the critics is unknown although the political necessity of having to have a Code in place by August 2025 suggest that the codes will be adopted as an interim measure; however, whether they are enough to win round its critics is not yet clear. The mood music to-date is not sounding positive from the AI provider side. While the Act states that the AI Office may “invite” providers of GPAI models to adhere to codes of practice, given that the AI office powers are considered draconian by the AI industry, it remains to be seen how many will voluntarily sign up to this GPAI Code.



Designed as a voluntary compliance mechanism, the Code plays a crucial role in the interim regulatory landscape leading up to the application of the AI Act's GPAI provider obligations from August 2025, and o offers practical guidance for AI providers on meeting specific obligations under the AI Act and for demonstrating early adherence to Articles 53 and 55; however it appears increasingly unlikely that the major stakeholders concerned with Fundamental Rights will support the adoption of the Code.

The Code is now subject to review for adequacy by EU Member States and the European Commission. If deemed appropriate, the Code will be endorsed and GPAI model providers will be able to use it to help demonstrate compliance with the EU AI Act. It is expected that the Code will be approved via an implementing act, conferring general validity across the EU. In practice, it is unlikely that the code will not be approved by the EU as this would be deeply embarrassing for the EU.

In fact, several prominent organizations that are likely to be considered providers of GPAI models have already stated their intention to sign onto the code and follow its measures as part of their AI governance program and this is largely because the Code has been significantly weakened following lobbying by Big AI. The anticipated template was not published alongside the new code.

The Code of Practice is primarily relevant for providers of GPAI models, such as – but not limited to – the well-known large-language models GPT (Open AI), Gemini, (Google) or the image generators Midjourney (Midjourney, Inc.) and DALL-E (Open AI). Specifically, the Safety and Security Chapter is relevant for providers of general-purpose AI models with systemic risk.

"Downstream providers", i.e. businesses implementing GPAI models, should familiarise themselves with the Code, too. The Code will likely have an impact on what developers of AI systems can expect and not expect from GPAI models, and influence negotiations of contracts with GPAI model providers.

A key change in the final text is that the level of detail required in technical documentation should be proportionate to the size of the model provider and is required to be updated regularly with documentation being kept for 10 years after market withdrawal

One of the key criticisms is that whilst Signatories are encouraged to disclose the Model Documentation (or parts thereof), there is no general obligation to publish it and this reflects the fact that the documentation is intended to be provided to the supervisory authority (on request) and to downstream providers of AI Systems, subject to certain confidentiality requirements.

Only if necessary to assess and/or mitigate systemic risks, will signatories have to publish anything and then only a summarised version of their Framework and Model Report(s) and following lobbying from the AI industry, GPAI providers are exempt from disclosing the amount of energy used to train a model

In addition, there are major concerns about the transparency requirements revealing trade secrets, confidential information and market-sensitive information. The transparency obligations expressly exclude the need to provide these in many cases, but even where there is no exclusion, it is highly unlikely that any AI Board will adhere to the code where it is required to disclose this information or any other information that could provide an advantage to a competitor. In the majority of cases, a meaningful provision of information to the public that could have an impact in share-prices or which revealed trade secrets, confidential information or market-sensitive information would be a breach of the obligations owed by the corporate officers and employees to the company and could open them to direct legal action. As a result, it can be expected that the transparency information will be poured over by both technical teams as well as lawyers ensuring that the code is technically complied so that there is a gloss of transparency whilst, in practice, not providing any meaningful transparency.

In the context of AI development, "trade secrets" and "confidential information" are not merely abstract legal terms; they represent the core intellectual property that gives an AI company its competitive edge. This includes training-data and data-curation methodologies including proprietary datasets, being vast, unique, and often privately curated datasets for which many millions of dollars have been spent on creation and the training and datasets are critical elements in determining competitivity.

The sheer effort, cost, and specialized knowledge involved in collecting, cleaning, annotating, and structuring these datasets is enormous and disclosing the specific sources, composition, or even the detailed methodologies for data acquisition (e.g., specific web scraping techniques, partnerships for licensed data, internal data generation processes) can reveal years of competitive investment, meaning that this cannot effectively be done within transparency obligations and this can be seen by the differing public approaches by Google and Anthropic which have been revealed in US Court disclosures.

In addition, the methodologies used to filter, normalize, and pre-process raw data to remove noise, bias, or sensitive information are often highly sophisticated and proprietary and their disclosure would be deemed to be disclosure of market-sensitive information as revealing these techniques could allow competitors to replicate a model's performance or avoid pitfalls that the original developer spent significant resources overcoming. It also includes model architecture and design principles and while some foundational AI architectures are publicly known, the specific modifications, layer configurations, parameter choices, and novel components and unique architectures developed by individual companies are often necessarily closely guarded secrets, yet it is these particular architectures and designs that need to be considered for proper transparency of operation.

Further concerns arise around the disclosure of the types of data used for training and the capabilities of a GPAI model as this can signal a company's strategic focus or future product directions, so for example a heavy investment in training data related to a specific industry (e.g., healthcare, finance) could indicate a strategic pivot or a new intended GPAI market entry aimed at that market, something that competitors could infer a company's investment levels, technological advancements, and market positioning based on the disclosed information, allowing them to adjust their own strategies.

There is also criticism that the European Commission's templates and measures outlined in the Codes of Practice focuses significantly on "major" or "large" datasets (e.g., those representing more than 5% of the total training data) and this threshold opens the door to artificially splitting large datasets into smaller subsets to avoid disclosure, thereby distorting the true picture of the training data and transparency whilst also raising concerns that significant portions of problematic data might remain undisclosed if they fall below this arbitrary threshold. Indeed many critics note that the Codes of Practice appears overly focused on copyright protection, potentially overlooking other crucial pre-processing steps, and the potential role of bad actors as well as failing to address regulation of matters such as the methods of anonymization or data filtering, which are vital for understanding bias mitigation and privacy preservation as well as providing significant and inappropriate exclusions for Open Source AI models.

In addition to these design choices are also the specific algorithms, hyperparameter tuning strategies, and optimization routines used during model training that are critical to achieving high performance and efficiency and which are highly proprietary and represent significant intellectual property and simply disclosing the exact number of layers, the type of attention mechanisms, or the specific loss functions used in a GPAI model together with other transparency-driven architectures could provide a blueprint for competitors to build a similar model without having to go through the same costly research and development cycles as appears, according to trade news reports, to have occurred with the development of DeepSeek.

Critical to true transparency is an understanding of the training regimes, the precise sequence of training steps, the duration of each phase, the specific hardware configurations, and the computational resources (e.g., FLOPs, GPU hours) consumed during training which are both price-sensitive information as well as critical and highly sensitive confidential information. Similarly to understand the true risk presented by a model as is necessary for safety and security assessments and which the transparency obligations were intended to highlight for external evaluators it is necessary to understand the fine-tuning and alignment techniques, but in many cases these are not even fully understood internally and even if they were, they are considered so proprietary that no disclosure would be permitted with the result that the risks arising from fine-tuning, reinforcement learning from human feedback (RLHF), or other alignment techniques that shape a model's behaviour, all critical to effective safety assessment, have had to be excluded to make the transparency codes acceptable top model providers.

Similarly, trust in regulatory bodies and data handling is a major issue both for the Transparency Code of Practice and the Safety and Security Code of Practice where AI developers are inherently cautious about sharing highly sensitive proprietary information with any external entity, including regulatory bodies like the AI Office or national competent authorities because of concerns over the risk of data breaches or leaks from regulatory databases, concerns over whether regulatory staff possess the technical expertise to handle and interpret highly complex AI-related trade secrets without inadvertently compromising them and whether regulatory staff with access to sensitive material may leave and go to work for competitors and the very real concern that even within the EU, there are significant variations in how different national authorities handle and protect confidential information.Whilst the impacts of these systems and the model's safety, performance, and ethical alignment are critical to transparency but are also proprietary and excluded from public disclosure and are unlikely under any circumstances to be disclosed to an AI office in any meaningful manner due to their sensitivity and the fact that the emergence of the AI Offices and national regulatory authorities have created a single source of stored critical data. As the recent UK Afghan data leak has shown, governments and government offices cannot be trusted to implement even the most basic of security safeguards despite the leak costing over £7Bn, and the leak was covered up for 2 years by senior ministers.

The costs of compromise of AI Offices if the aforesaid information was provided and compromised would run into trillions of dollars and therefore the concept of transparency in the EU AI act Code of Conduct is entirely illusory and the loss of the Afghan data and other material data leaks from elsewhere in the EU are likely to be used as examples of why transparency and critical safety and security data cannot be disclosed, regardless of consequences, under the Act. The parallel concern is also that if the regulatory authorities seek to force disclosure then the AI companies will simply exclude use of their AI models in Europe and in relation to European user data so that they do not fall within the remit of the EU Ai act, with the consequential significant economic damage that this would do to the EU.

The same difficulties of lack of effective transparency also apply to internal benchmarking data where proprietary internal benchmarks and evaluation datasets to rigorously test their models for accuracy, robustness, and safety before public release have been developed but where these internal evaluations, especially concerning limitations or failure modes, are highly confidential and trade secrets protected under the Act from disclosure and which in the hands of bad actors would enable guard-rails and safeguards to be compromised and where it is not in the public interest that the specific methodologies used to identify and mitigate biases, detect adversarial attacks, or measure the model's "truthfulness" are disclosed. Revealing the exact performance scores on specific internal safety benchmarks could give competitors insights into a model's weaknesses, allowing them to focus their own development efforts or even exploit those weaknesses.

As the EU AI Act and its Code of Practice expressly acknowledge the need to protect trade secrets and state[2] that the transparency obligations should be applied "without prejudice to the protection of intellectual property rights or trade secrets" and the Codes of Practice talks about balancing transparency with the need to protect confidential information, trade secrets etc. Despite these express exclusions designed to protect AI providers, there remain concerns about ambiguity in language[3], such as core issue lies in the interpretation of what constitutes "strictly necessary" disclosure versus what falls under "trade secret" protection.

For example, what a regulator deems necessary for oversight, a company might view as a critical competitive asset falling within trade secret protection from disclosure. Similar concerns exist over the term "*sufficiently detailed*" which is inherently vague and it is highly questionable whether the high-level summary will be sufficient for evaluators or downstream providers to be able to effectively understand the models, with an

[2] Article 53(2) of the Act

[3] Ambiguity in language is not uncommon in EU projects and initiatives, because although the language used for most projects in English, the majority of participants do not have this as a first language so documents are translated with loss of precision and changes of meanings as well as suffering from a lack of detailed analysis of the final text with the result that final language is inherently ambiguous. Indeed, there can be ambiguity simply from differing language uses between US English and UK English.

inherent technical infeasibility of being transparent for models which have granular itemization of trillions of data points. What is sufficiently detailed from the point of view of the provider may be different from what is sufficiently detailed from the point of view of the downstream integrator and in turn what is sufficiently detailed from the point of view of the downstream integrator may be different from the what is sufficiently detailed from the point of view of the regulator or copyright rightsholders. As a result, there is significant legal uncertainty in the codes where companies are required by Company law and trade secret law, as well as third party contracts, to err on the side of caution, leading to less transparency, being faced on one hand with a threat that they are non-compliant in transparency according to regulatory authorities, whilst being faced with the threat of legal action for breach of secrecy obligations to the Company by shareholders, breach of disclosure of market sensitive data by financial regulators and breach of confidentiality contracts by third parties.

Without clear, measurable criteria for what constitutes "sufficiently detailed," the transparency obligation risks becoming a mere formality rather than a substantive tool for accountability. This ambiguity can lead to differing interpretations and potential disputes, undermining the very goal of transparency.

Another example is the use of terms such as "Sufficiently Detailed Summary" because even if specific URLs or information sources are not required to be disclosed, the demand for a "sufficiently detailed summary" of training data can still be problematic because it can provide sufficient disclosure that rights-holders can seek further disclosures, litigators can refer to redacted documents disclosed under the EU AI act to claim that the document has been disclosed and therefore the entire document without redactions should be disclosed to in Court litigation and even if the summary requires listing "major" datasets, categories of data sources, or general data acquisition methods, it can still provide valuable insights to competitors and create speculative actions from ambulance-chasing rights-holder lawyers. The fear is that even aggregated or high-level information, when combined with publicly available data or other competitive intelligence, could be pieced together to reverse-engineer aspects of the model or training process in the same way that AI has allowed previously anonymous medical data to be paired with other data in order to identify particular patients.

Similar challenges lie in the Codes where defining the "strictly necessary information" for regulators versus what needs to be disclosed to the public or downstream users. This is likely to be a major source of limitation between AI providers and EU AI Officers and national regulators.

In addition, SMEs and SMCs will find the comprehensive documentation and information-sharing requirements disproportionately burdensome, potentially hindering their ability to innovate and compete, as the practical implementation of disclosure obligations, even under the Codes could still pose significant resource challenges.

There are also concerns about the effect that the EU AI Act may have on Innovation in the EU given the chilling effect of *fear* of inadvertently disclosing trade secrets or facing legal challenges related to transparency or Safety and Security as AI companies become more “EU Sensitive” in relation to their research and development, particularly in areas that require vast and diverse datasets, to minimize perceived disclosure risks via Europe. Leading to a situation where European AI companies are less willing to push the boundaries of AI development, especially for frontier models, compared to competitors in jurisdictions with less stringent transparency requirements and less stringent safety and security disclosure obligations. There are also material concerns that leaks of Transparency, Safety and Security information from AI Offices could lead to downstream breaches of guardrails and safeguards with the result that GPAISRs are retrained by bad actors and released with catastrophic and techsistential[4] risks for which there will be liability and compensation actions against the provider of the GPAI and denials of liability from AI Offices. In light of this, a number of AI providers have stated that the requirement to answer specific questions or requests for additional detail from AI Offices in additional to the information that AI providers are happy to provide publicly would only be forthcoming upon a) an audit by the AI provider of the AI Office staff and security procedures as well as full government backed indemnities because **the c**ompanies that have already invested heavily in developing

[4] Existential risks from the technology

advanced GPAI models and rely on their proprietary data and methodologies as a key differentiator, believe that forced transparency could erode this first-mover advantage.

Given these concerns, it is highly likely that AI providers, especially those with significant proprietary assets, will approach the Codes of Practice's transparency requirements as well as the Security and Secrecy Code of Practice with caution and potentially resistance. They will use:
a) **Strategic Interpretation** where they and their lawyers will interpret "sufficiently detailed" in the narrowest possible way that they believe still meets the formal requirement, while minimizing the disclosure of truly sensitive information, and will take legal action to protect that interpretation if threatened with sanctions by the regulators;
b) **Extensive Lobbying** to ensure flexibility in interpretation and to ensure that future iterations of the Codes of Practice or supplementary guidelines provide maximum flexibility and clear internal safe harbours for proprietary information where regulators have access only to information in the confines of the AI providers offices and under strict view-only no-copy conditions and where the regulatory and evaluation individuals with access are prohibited from working for competitors for a particular number of years[5];
c) **Focus on Compliance with Core Act** where AI companies will prioritize strict compliance with the legally binding aspects of the AI Act, ignoring adherence to the voluntary Codes of Practice or possibly being very selective about these , focusing on areas that align with existing internal practices or where the disclosure risk is minimal, with the effect that the EU AI Act will need to be constantly updated or will become rapidly outdated and irrelevant;
d) **Risk-Benefit European Analysis** where each AI company will conduct its own risk-benefit analysis and if the perceived risk of trade secret disclosure outweighs the benefits of adhering to the voluntary Codes of Practice (e.g., reduced administrative burden, legal certainty), or operating in Europe, they might either exclude Europe or choose to demonstrate compliance through other means;
e) Legal Challenges For the above reasons, where regulatory bodies push for disclosure levels that AI companies deem to be a threat to or a violation of their trade secrets, legal challenges are very highly probable, testing the boundaries of the Act's provisions in court, whilst also ensuring that previous compliance with AI office requests are not used against them, with the result that each and any request for disclosure is met with very expensive and elongated litigation.

In essence, the tension between the EU's desire for transparent and trustworthy AI via the Codes of Practice and ALTAI, and the parallel tension with the Safety and Security Code of Practice and the industry's need to protect its intellectual property is a fundamental challenge that the Codes of Practice have come nowhere close to solving. While the AI Act attempts to strike a balance, the practical implementation of Transparency and Safety and Security obligations, particularly within the voluntary Code of Practice, will continue to be a battleground where companies prioritize safeguarding their competitive advantage.

The effectiveness of these transparency measures will ultimately depend on the development of clear, enforceable guidelines that genuinely balance these competing interests without stifling innovation.

Whilst the main battleground will be between transparent and trustworthy AI via the Codes of Practice and ALTAI, and the parallel tension with the Safety and Security Code of Practice and the industry's need to protect its proprietary and intellectual property, the Copyright aspect is likely to take a back seat, especially given the number of cases working their way through European and US Courts. The copyright holders and civil society groups will continue to demand much greater granularity (e.g., specific titles, authors, or URLs) to effectively verify if their works have been used and to pursue infringement claims and to develop increasingly ill-founded claims that AI generated works somehow infringe their rights.

In this respect, their claims mirror the advocates for the equine economy in the early 20th century who saw the noisy, polluting, and often dangerous automobile as an existential threat, just as the music industry tried to hang onto CDs in the face of internet music, as AI, particularly generative AI, LLM and LRMs challenge the

[5] (probably 5 or more years by which time the proprietary information will be less valuable This will cause significant engagement and loss of work opportunity indemnity costs for regulators and evaluators.

very entrenched and human centric protection of creativity, protecting an increasingly unsustainable economic model for authors, artists, and musicians to control any form of threat to their entrenched protectionism. As AI models, trained on vast datasets of copyrighted material, can generate text, images, and music at an unprecedented speed and scale, often mimicking human styles, the rights-holders have correctly perceived this as an existential threat, with lawsuits demanding ever stricter regulations whilst the Courts continue to agree that direct human authorship is the sole basis for protection and only human generated works, not copies in similar styles are capable of protection. However, just as the internal combustion engine ultimately reshaped society, the transformative power of AI is likely to necessitate a fundamental re-evaluation and adaptation of copyright law and the Code of Conduct recognises this, creating a complex coexistence, where AI becomes an increasingly powerful tool but one without any protection in its output and where the obligation for appropriate measures to prevent training on works have to be implemented by the artists, reflecting the previous obligation of the artist to sue forgers for copying their works.

The Codes of Practice's voluntary nature is a fundamental point of contention, particularly for stakeholders seeking stronger compliance and accountability because **w**hile adherence to the Code is intended to help providers demonstrate compliance with the AI Act, it does not automatically grant a legal presumption of conformity, meaning that there is materially reduced incentive to adhere to the Codes of Practice as even if a provider follows these, they could still be found non-compliant with the Act itself. This weakens the incentive for full and rigorous adherence, as providers are likely to choose to demonstrate compliance through other, potentially much less transparent, methodologies. Similarly for those GPAI providers who choose not to sign or fully implement the commitments under the Codes of Practice, the direct enforcement mechanisms are less clear because whilst the AI Act itself carries penalties, the voluntary nature of the Codes of Practice means that it necessarily lacks direct enforcement power and, upon an unacceptable request for detail from the AI Office, AI providers can simply announce that they are not strictly following the Codes, providing other measures to comply with the AI Act, and thereby that they do not need to answer the request for details, placing a huge, expensive and possibly practically impossible burden on AI officers of having to otherwise determine how the provide is not complying with the AI Act, thus creating an uneven playing field between Code followers and others and allow some providers to avoid the spirit of the transparency and the safety and security obligations. There is further criticism that the voluntary nature of the Codes of Practice suggested that " documentation" would primarily be provided to the AI Office[6] and critics have argued that this approach falls short of true public transparency, failing to convince the EU that information about training data and model capabilities is of public interest.

The Green lobby have complained that the removal of requirements in the Codes to provide energy consumption data is inappropriate but the difficulty of providing meaningful and effectively accurate data has mandated this, as highlighted by some AI providers, due to differing energy consumption costs, it is open to AI providers to choose their lowest consumption data criteria and to manipulate the answer to reflect only this particular consumption model. Other critics argue that disclosure should also explicitly cover water consumption and location-based factors affecting cooling needs, but this has been resisted as overly burdensome.

Overall, the criticisms of the EU AI Act Code of Practice highlight the inherent difficulties in regulating a fast-moving, complex technology like AI and while the Codes of Practice is a necessary step to operationalize the Act's principles and to be changed in light of advances, its voluntary nature, the limited nature of the Codes and the ongoing ambiguities in key definitions will represent significant challenges.

[6] and national competent authorities upon request, although the lack of definition of national competent authorities is likely to result in significant litigation to establish the authority of law to make such a request by those claiming to be a national competent authority.

1. Transparency Code of Practice

Commitment: GPAI model providers must provide a standardized and comprehensive record that can be accessed by relevant parties and the key measures to achieve this.

The third draft of the EU AI Act's Code of Practice for General-Purpose AI (GPAI) models, particularly its transparency section, was intended to ensure that providers of these foundational models gave sufficient information about their models and how they were trained as this was considered crucial for downstream AI system developers, the EU AI Office, and potentially copyright holders, to understand the models' capabilities, limitations, and compliance with legal obligations and to comply with their transparency obligations under Article 53(1), points (a) and (b), and the corresponding Annexes XI and XII of the AI Act.

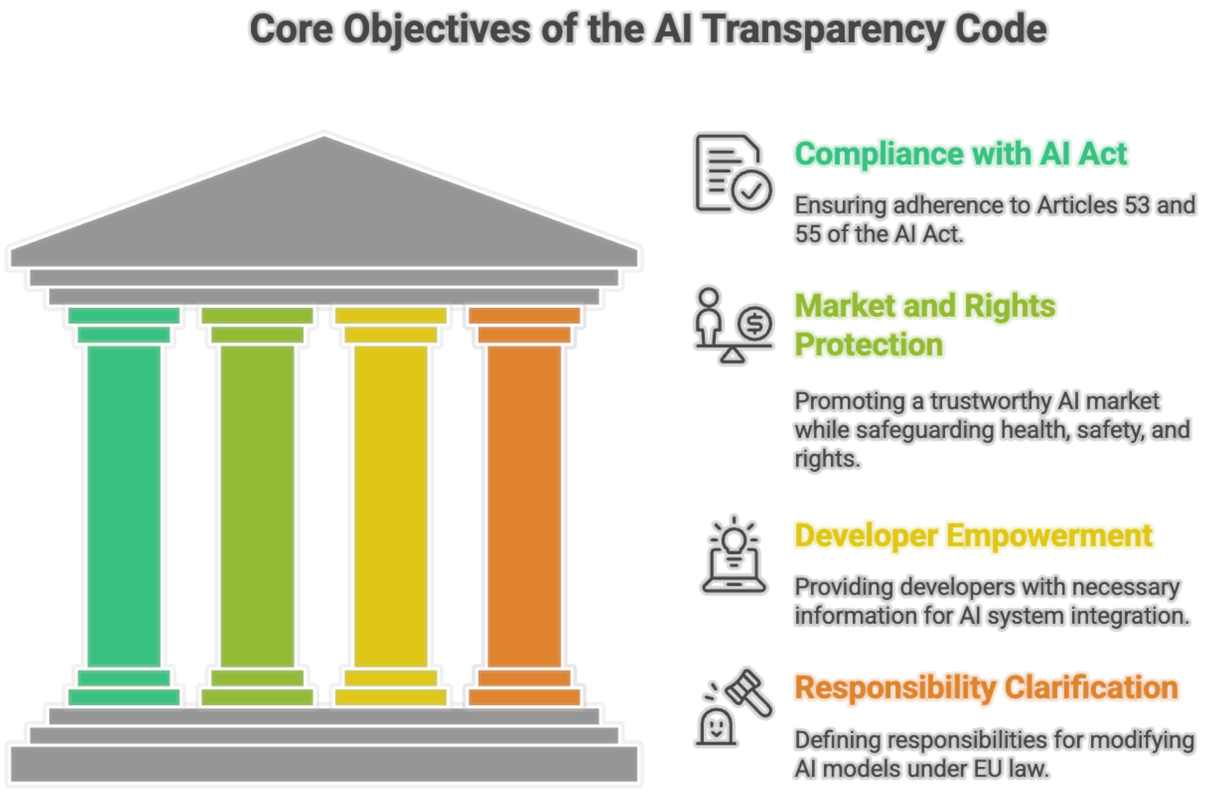
As a result of the extensive lobbying of the copyright holders as well as the activists in the green addenda, coupled with counter-lobbying of the AI industry, the final version of the Transparency codes is an attempt to compromise, although in practice the Transparency Codes now does little to set out a comprehensive transparency obligation for AI systems and entirely remove the need within the Codes to disclose energy requirements. The compromises arose largely because the EU Experts accepted that it is often the case that even the developers of the AI systems are unable to explain how the GPAI models work in practice and that energy consumption is both price sensitive and also impossible to provide in a practically useful manner given differing energy costs throughout the world, so although earlier drafts expressly allowed environmental protection via energy in training questions, these do not exist in the final version.

The Model Documentation Form indicates for each item whether the information is intended for downstream providers, the AI Office or national competent authorities.
a) For the AI Office or national competent authorities, a specific request must be made from the relevant authority for that information and that request must state the legal basis and purpose of the request. Compliance is only necessary if the requesting party has established legal basis and that the information

requested is strictly necessary for the AI Office to fulfil its tasks under the AI Act, in particular to assess compliance of providers high-risk AI systems built on general-purpose AI models where the provider of the system is different from the provider of the model.

In particular, under Article 78 AI Act, the recipients of any of the information contained in the Model Documentation Form is obliged to respect the confidentiality of the information obtained, in particular intellectual property rights and confidential business information or trade secrets, and to put in place adequate and effective cybersecurity measures to protect the security and confidentiality of the information obtained and it can be expected, given the enormous damages that would flow from a breach of the secrecy obligations, that providers of the information will require the requesting AI Office to show how they will fulfil that obligation prior to disclosure. In the interim, it can be expected that information will be written in such as manner that it does not disclose intellectual property rights and confidential business information or trade secrets.

The aim of the Code is
a) to demonstrate compliance with the obligations provided for in Articles 53 and 55 AI Act and
b) to “improve the functioning of the internal market”, “to promote the uptake of human-centric and trustworthy artificial intelligence (“AI”)”, and to ensure “a high level of protection of health, safety, and fundamental rights” and to protect against harmful effects of AI in the Union, and
c) to ensure that developers of downstream AI systems have adequate information and a good understanding of the models and their capabilities to properly integrate them as well as meet AI Act obligations as typically GPAI (general-purpose AI) models may form the basis for a range of downstream AI systems,
d) to clarify that when a person, public authority, agency or other body modifies the GPAI model by fine tuning it, that entity becomes the provider of the modified model and is subject to the EU AI Law obligations for providers, but that responsibility should be limited to that modification or fine-tuning.



Core Objectives of the AI Transparency Code

Compliance with AI Act

Ensuring adherence to Articles 53 and 55 of the AI Act.

Market and Rights Protection

Promoting a trustworthy AI market while safeguarding health, safety, and rights.

Developer Empowerment

Providing developers with necessary information for AI system integration.

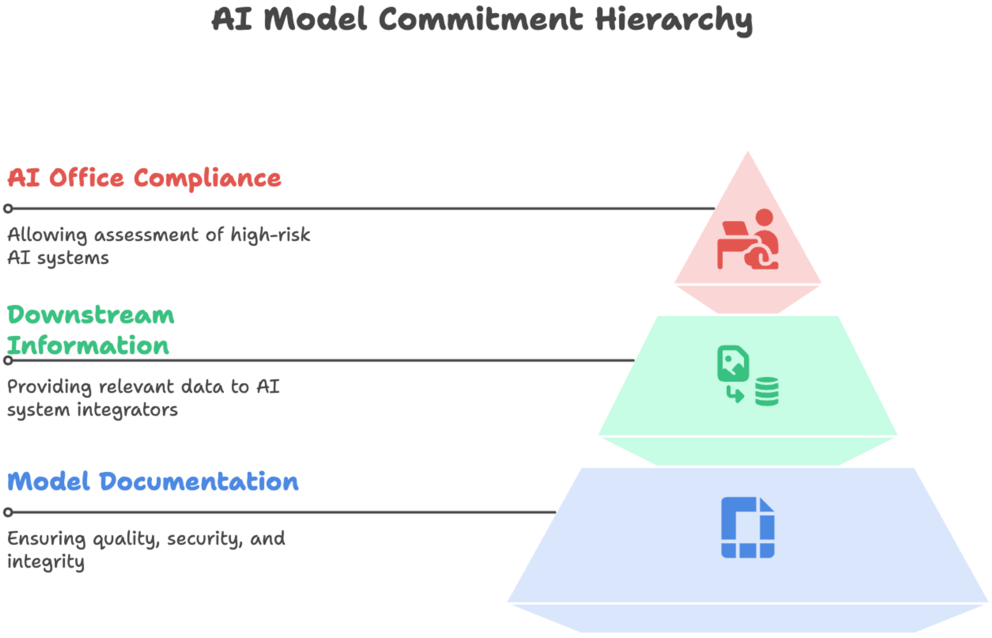
Responsibility Clarification

Defining responsibilities for modifying AI models under EU law.

Although earlier drafts had an express presumption that compliance with the code was compliance with the obligations in Articles 53 and 55 AI Act, this is no longer the case in the final version and adherence to the Code no longer constitute conclusive evidence of compliance with those obligations.

The Commitments are:
a) a commitment to draw up and keeping up-to-date model documentation, ensuring quality, security, and integrity of the documented information;
b) a commitment to provide relevant information to providers of AI systems who intend to integrate the general-purpose AI model into their AI systems (‘downstream providers’), and
c) to provide information to the AI Office upon request (subject to being strictly necessary for the exercise of their supervisory tasks) and to allow the AI officer to assess the compliance of a high-risk AI system built on a GPAI model where the provider of the system is different from the provider of the model;

Note: GPAI models released under a free and open-source license are excluded unless the model is a general-purpose AI model with systemic risk.


AI Model Commitment Hierarchy
AI Office Compliance
Allowing assessment of high-risk AI systems
Downstream Information
Providing relevant data to AI system integrators
Model Documentation
Ensuring quality, security, and integrity

I.1 Model Documentation

Drawing up and keeping up-to-date Model Documentation (Measure I.1.1):
Providers are required to prepare a document titled "Information and Documentation about the General-Purpose AI Model" (referred to as "Model Documentation") when they place a GPAI model on the market.

This documentation must contain all the information specified in a standardized "Model Documentation Form." This form is intended to ensure consistency and comparability across different models.
The documentation needs to be updated to reflect any relevant changes to the model.

Providers must keep previous versions of the Model Documentation for a period of 10 years after the model has been placed on the market. This ensures a historical record for oversight and accountability.

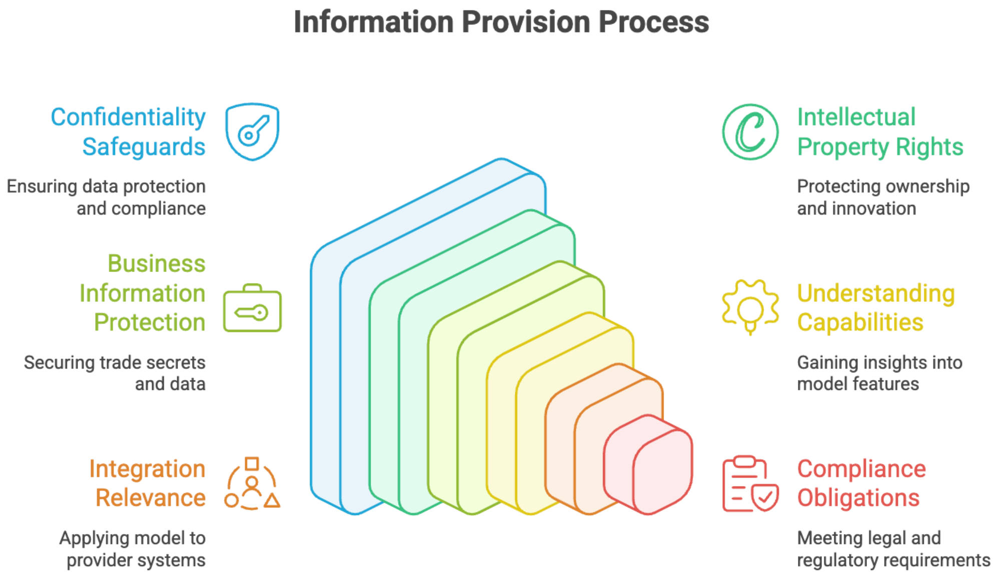
Crucially, the documentation must no longer report information on computational resources and energy consumption in a way that is consistent with any delegated acts (more specific rules) adopted under Article 53(5) of the AI Act which aims for comparable and verifiable data on the environmental impact of training these models. This is a reflection both of lobbying and a realisation that the disclosure requires disproportionate effort, is largely meaningless in light of differing world energy pricing and is share price sensitive.

Measure 1.2 Providing relevant information

When placing a general-purpose AI model on the market, it will be necessary for providers to publicly disclose via their website, or via other appropriate means if they do not have a website, contact information for the AI Office and downstream providers to request access to the relevant information contained in the Model Documentation, or other necessary information.

Certain information is only to be provided to the AI office, upon a request from the AI Office and this includes the information in the model form as AI office information as well as any additional information *necessary* for the AI Office to fulfil its tasks under the AI Act[7], in particular to assess compliance of high-risk AI systems built on general-purpose AI models where the provider of the system is different from the provider of the model.

Although there is an obligation to provide to downstream providers the information contained in the most up-to-date Model Documentation that is intended for downstream providers, this is subject to the confidentiality safeguards and other conditions provided for under Articles 53(7) and 78 AI Act and in particular the need to observe and protect intellectual property rights and confidential business information or trade secrets in accordance with Union and national law. If downstream providers are able to satisfy these safeguards, then they should be provided with additional information as
a) necessary to enable them to have a good understanding of the capabilities and limitations of the general-purpose AI model
b) relevant for its integration into the downstream providers' AI system and
c) necessary to enable those downstream providers to comply with their obligations pursuant to the AI Act.


Information Provision Process
Confidentiality Safeguards
Ensuring data protection and compliance
Business Information Protection
Securing trade secrets and data
Integration Relevance
Applying model to provider systems
Intellectual Property Rights
Protecting ownership and innovation
Understanding Capabilities
Gaining insights into model features
Compliance Obligations
Meeting legal and regulatory requirements

Although earlier drafts required this on a reasonable timeframe, this is not no later than 14 days of receiving the request save for exceptional circumstances.

Although there is no obligation to do so, all Providers are "encouraged" to consider whether the documented

[7] or for national competent authorities to exercise their supervisory tasks under the AI Act

information can be disclosed, in whole or in part, to the public to promote public transparency, even if in a summarised form.

Although there has been extensive lobbying for explicit and detailed training data details to be provided, the final version of the code only requires that information used as part of the training content is only provided in summarized forms although those summaries[8] must be made publicly available under Article 53(1), point (d), AI Act.

Some parties had sought to argue that transparency requirements from the AI Act are intrinsically linked to transparencies in Training Data (Article 53(1)(d) AI Act) which is a critical transparency obligation mandated by the AI Act itself, which states that GPAI model providers must "draw up and make publicly available a sufficiently detailed summary about the content used for training of the general-purpose AI model, according to a template provided by the AI Office."

This summary is expected to be comprehensive but not overly technical, balancing transparency with the protection of trade secrets. As a result of these provisions, and US litigation, Anthropic revealed that it spent millions of dollars physically scanning print books to build the AI system Claude, cutting millions of print books from their bindings, scanned them into digital files, and throwing away the originals solely for the purpose of training AI, something designed to replicate Google's legally successful book digitization approach—the same scanning operation that survived copyright challenges and established key fair use precedents[9].

Anthropic destroyed millions of print books to build its AI models

This destructive scanning operation qualifies as fair use because Anthropic had legally purchased the books first, destroyed each print copy after scanning, and kept the digital files internally rather than distributing them, a process known as “format conversion”. Anthropic has previously admitted that it originally used pirated libraries for training and it is considered that that early training on pirated libraries was illicit, but the retaining on format converted books was legal under the first sale doctrine in the USA. The same doctrine would apply under EU law where the doctrine of Exhaustion of Rights applies upon the first sale.

While the AI Act mandates a "sufficiently detailed summary" of training data, the exact level of detail remains a point of contention as rightsholder groups often advocate for more granular disclosure to effectively enforce their rights and the code has, as is typical of EU legislation, created a half-way house where there is disclosure without requiring the disclosure of individual copyrighted works. Given the emerging caselaw from various jurisdictions that it is for the Artist (and their agents) to provide restrictions on the use of their work, whether by paywalls, subscriber-walls or machine readable exclusions and that if this is not done when training of freely available internet data is not copyright infringement, this is a legitimate and cohesive step.

Measure 1.3 Ensuring quality, integrity, and security of information

Providers must also ensure that the documented information is controlled for quality and integrity, retained

[8] See a template to be provided by the AI Office.
[9] The Google Books project largely used a patented non-destructive camera process to scan millions of books borrowed from libraries and later returned)

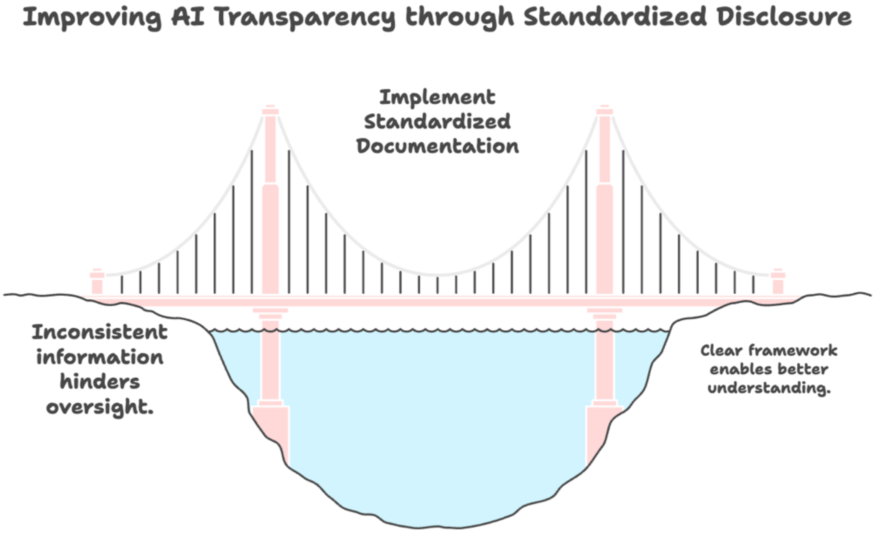
as evidence of compliance with obligations in the AI Act, and protected from unintended alterations.

In the context of drawing-up, updating, and controlling the quality and security of the information and records, the obligation is to follow the established protocols and technical standards of beast practice.

Given the protection that the disclosure under transparency obligations is subject to restrictions on confidentiality and trade secrets, the value of the transparency Code of Practice will depend upon the extent to which providers and downstream providers are willing to disclose information within the "standardized Model Documentation Form" and how much effort is put into appearing to comply with the disclosure obligation whilst not actually disclosing critically sensitive or price-sensitive information, and the Codes recognise this as a positive step towards ensuring consistent and comparable information across different GPAI models, whilst necessarily needing to strike a balance between providing sufficient transparency and protecting the valuable trade secrets of AI developers, something that is a perpetual tension in AI regulation.

There is particular concern in parts of the AI Ethics and Safety industry that there is lesser transparency for open-source GPAI models, where the documentation and transparency obligations are reduced even further unless they are classified as systemic-risk models, and the assessment of what is a systemic-risk model again, mostly lies with the developer, this introducing significant legal complexity and hurdles in the place of any attempt to claim a breach of the transparency codes.

There is also material concern at the very limited nature of the Model Documentation Form, given its very limited scope couple with the very low recommended number of words in each section of the form at critical parts of disclosure. In essence, the aim of the transparency section in the draft of the EU AI Act's Code of Practice for GPAI models was to create a clear framework for information disclosure, allowing for better oversight by authorities and greater understanding by those who use and are affected by these powerful AI models, although it is unlikely that this has been achieved in practice.


Improving AI Transparency through Standardized Disclosure
Implement Standardized Documentation
Inconsistent information hinders oversight.
Clear framework enables better understanding.

Model Documentation Form

This Form includes all the information to be documented as part of Measure 1.1 of the Transparency Chapter of the Code of Practice. Crosses on the right indicate whether the information documented is intended for the AI Office (AIO), national competent authorities (NCAs) or downstream providers (DPs), namely providers of AI systems who intend to integrate the general-purpose AI model into their AI systems. Whilst information intended for DPs should be made available to them proactively, information intended for the AIO or NCAs is only to be made available following a request from the AIO, either ex officio or based on a request to the AIO from NCAs. Such requests will state the legal basis and purpose of the request and will concern only items from the Form strictly necessary for the AIO to fulfil its tasks under the AI Act at the time of the request, or for NCAs to exercise their supervisory tasks under the AI Act at the time of the request, in particular to assess compliance of high-risk AI systems built on general-purpose AI models where the provider of the system is different from the provider of the model.

Any elements of information from the Model Documentation Form shared with the AIO and NCAs shall be treated in accordance with the confidentiality obligations and trade secret protections set out in Article 78 AI Act.

Date this document was last updated: Click or tap to enter a date. **Document version number:** Click or tap here to enter text.

General information		AIO	NCAs	DPs
Legal name for the model provider:	Click here to add text.	☒	☒	☒
Model name:	The unique identifier for the model (e.g. Llama 3.1-405B), including the identifier for the collection of models where applicable, and a list of the names of the publicly available versions of the concerned model covered by the Model Documentation.	☒	☒	☒
Model authenticity:	Evidence that establishes the provenance and authenticity of the model (e.g. a secure hash if binaries are distributed, or the URL endpoint in the case of a service), where available.	☒	☒	☐
Release date:	Click or tap to enter a date. Date when the model was first released through any distribution channel.	☒	☒	☒
Union market release date:	Click or tap to enter a date. Date when the model was placed on the Union market.	☒	☒	☒
Model dependencies:	If the model is the result of a modification or fine-tuning of one or more general-purpose AI models previously placed on the market, list the model name(s) (and relevant version(s) if more than one version has been placed on the market) of those model(s). Otherwise write 'N/A'.	☒	☒	☒

Model properties		AIO	NCAs	DPs
Model architecture:	A general description of the model architecture, e.g. a transformer architecture. *[Recommended 20 words]*	☒	☒	☒
Design specifications of the model:	A general description of the key design specifications of the model, including rationale and assumptions made, to provide basic insight into how the model was designed. *[Recommended 100 words]* If any other please specify:	☒	☒	☐
Input modalities:	☐Text ☐Images ☐Audio ☐Video If any other please specify:	☒	☒	☒
For each selected modality please include maximum input size or write 'N/A' if not defined	Maximum size: … Maximum size: … Maximum size: … Maximum size: … Maximum size: …	☒	☐	☒
Output modalities:	☐Text ☐Images ☐Audio ☐Video If any other please specify:	☒	☒	☒
For each selected modality please include maximum output size or write 'N/A' if not defined	Maximum size: … Maximum size: … Maximum size: … Maximum size: … Maximum size: …	☐	☐	☒
Total model size:	The total number of parameters of the model, recorded with at least two significant figures, e.g. 7.3*10^10 parameters.	☒	☐	☐
The range within which the total number of parameters falls.	☐1—500M ☐500M—5B ☐ 5B—15B ☐15B—50B ☐50B—100B ☐100B—500B ☐500B—1T ☐>1T	☐	☒	☒

Methods of distribution and licenses		AIO	NCAs	DPs
Distribution channels:	A list of the methods of distribution (e.g. enterprise or subscription-based access through existing software suites or enterprise-specific solutions; public or subscription-based access through an API; public or proprietary access through integrated development environments, device-specific applications or firmware, open-source repositories) through which the model has been made available for distribution or use in the Union market. For each listed method of distribution, please include a link to information about how the model can be accessed, where available, and the level of model access (e.g. weights-level access, black-box access).	☒	☒	☐

	A list of the methods of distribution (e.g. enterprise or subscription-based access through existing software suites or enterprise-specific solutions; public or subscription-based access through an API; public or proprietary access through integrated development environments, device-specific applications or firmware, open-source repositories) through which the model can be made available to downstream providers.	□	□	☒
License:	A link to model license(s) (otherwise provide a copy of the license(s) upon a request from the AIO pursuant to Article 91) or indicate that no model license exists.	☒	☒	□
	The type or category of licence(s) under which the model can be made available to downstream providers such as free and open source licences where models can be openly shared and providers can freely access, use, modify and redistribute them or modified versions thereof; less permissive licenses that impose certain restrictions on the use (e.g. to ensure ethical use), or proprietary licences that restrict access to the model's source code and impose limitations on usage, distribution, and modification. In the absence of a license, describe how access to the model is provided for downstream use, such as through terms of service.	□	□	☒
	A list of additional assets (e.g. training data, data processing code, model training code, model inference code, model evaluation code), if any, that are made available with a description of how each can be accessed and what licenses, if any, relate to their use.	☒	□	☒
	Use	**AIO**	**NCAs**	**DPs**
Acceptable Use Policy:	Provide a link to the acceptable use policy applicable (or attach a copy to this document) or indicate that none exists.	☒	☒	☒
Intended uses:	A description of either (i) the uses that are intended by the provider (e.g. productivity enhancement, translation, creative content generation, data analysis, data visualisation, programming assistance, scheduling, customer support, variety of natural language tasks, etc..) or (ii) the uses that are restricted and/or prohibited by the provider (beyond those prohibited by EU or international law, including Article 5 AI Act), in both cases as specified in the information supplied by the provider in the instructions for use, terms and conditions, promotional or sales materials and statements, as well as in the technical documentation. If specifying (i) or (ii) is incompatible with the nature of the license under which the model is provided, then 'N/A' can be entered. *[Recommended 200 words]*	☒	☒	☒
Type and nature of AI systems in which the general-purpose AI model can be integrated:	A list or description of either (i) the type and nature of AI systems into which the general-purpose AI model can be integrated or (ii) the type and nature of AI systems into which the general-purpose AI model should not be integrated. Examples may include autonomous systems, conversational assistants, decision support systems, creative AI systems, predictive systems, cybersecurity, surveillance, or human-AI collaboration. *[Recommended up to 300 words]*	☒	☒	☒
Technical means for model integration:	A general description of the technical means (e.g. instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems. *[Recommended 100 words]*	□	□	☒
Required hardware:	A description of any hardware, including the version, required to use the model, where applicable. If not applicable (e.g. model offered via an API), 'N/A' should be entered. *[Recommended 100 words]*	□	□	☒
Required software:	A description of any software, including the version, required to use the model where applicable. If not applicable, 'N/A' should be entered. *[Recommended 100 words]*	□	□	☒
	Training process	**AIO**	**NCAs**	**DPs**
Design specifications of the training process:	A general description of the main steps or stages involved in the training process, including training methodologies and techniques, the key design choices, assumptions made and what the model is designed to optimise for, and the relevance of different parameters, as applicable. For example, "the model is initialized with randomly selected weights and optimised using gradient-based optimization via the Adam optimizer in two steps. First, the model is trained to predict the next word on a large pretraining corpus using the cross-entropy loss, passing over the data for a single epoch. Second, the model is post-trained on a dataset of human preferences for 10 epochs to align the model with human values and make it more useful in responding to user prompts". *[Recommended 400 words]*	☒	☒	□
Decision rationale:	A description of how and why key design choices were made in model training. *[Recommended 200 words]*	☒	☒	□
	Information on the data used for training, testing, and validation	**AIO**	**NCAs**	**DPs**
Data type/modality: *Select all that apply.*	□Text □Images □Audio □Video If any other please specify:	☒	☒	☒
Data provenance: *Select all that apply* For definitions of each listed category, see the Template for the Public Summary of the Training Content of General-Purpose AI models provided by the AI Office	□Web crawling □Private non-publicly available datasets obtained from third parties □User data □Publicly available datasets □Data collected through other means □Synthetic data that is not publicly accessible (when created directly by or on behalf of the provider) If any other please specify:	☒	☒	☒

How data was obtained and selected:	A description of the methods used to obtain and select training, testing, and validation data, including methods and resources used to annotate data, and models and methods used to generate synthetic data where applicable. For data previously obtained from third parties, a description of how the provider obtained the rights to the data if not already disclosed in the public summary of training data published in accordance with Article 53(1), point (d). *[Recommended 300 words]*	☒	☒	□
Number of data points:	The size (in number of data points) of the training, testing, and validation data respectively, together with the definition of the unit of data points (e.g. tokens or documents, images, hours of video or frames), recorded with at least one significant figure (e.g. $3x10^{13}$ tokens).	□	☒	□
	The size (in number of data points) of the training, testing, and validation data respectively, together with the definition of the unit of data points (e.g. tokens or documents, images, hours of video or frames), recorded with at least two significant figures (e.g. $1.5x10^{13}$ tokens).	☒	□	□
Scope and main characteristics:	A general description of the scope and main characteristics of the training, testing and validation data, such as domain (e.g. healthcare, science, law,...), geography (e.g. global, restricted to a certain region,...), language, modality coverage, where applicable. *[Recommended 200 words]*	☒	☒	□
Data curation methodologies:	General description of the data processing involved in transforming the acquired data into training, testing, and validation data for the model, such as cleaning (e.g. filtering out irrelevant content such as advertisements), normalisation (e.g. tokenizing), augmentation (e.g. back-translation). *[Recommended 300 words]*	☒	☒	☒
Measures to detect unsuitability of data sources:	A description of any methods implemented in data acquisition or processing, if any, to detect the presence of unsuitable data sources considering the model's intended uses, including but not limited to illegal content, child sexual abuse material (CSAM), non-consensual intimate imagery (NCII), and personal data leading to its unlawful processing. *[Recommended 400 words]*	☒	☒	□
Measures to detect identifiable biases:	A description of any methods implemented in data acquisition or processing, if any, to address the prevalence of identifiable biases in the training data. *[Recommended 200 words]*	☒	☒	□
Computational resources (during training)		**AIO**	**NCAs**	**DPs**
Training time:	A description of what period is being measured along with the range that its duration falls under, within the following ranges: less than 1 month, 1—3 months, 3—6 months, more than 6 months.	□	☒	□
	A description of what period is being measured along with the duration in wall clock days (e.g. $9x10^{1}$ days) and in hardware days (e.g. $4x10^{5}$ Nvidia A100 days and $2x10^{5}$ Nvidia H100 days), both recorded with at least one significant figure.	☒	□	□
Amount of computation used for training:	Measured or estimated amount of computation used for training, reported in floating point operations and recorded up to its order of magnitude (e.g. 10^{24} floating point operations).	□	☒	□
	Measured or estimated amount of computation used for training, reported in computational operations and recorded with at least two significant figures (e.g. $2.4x10^{25}$ floating point operations).	☒	□	□
Measurement methodology:	In the absence of a delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies, describe the methodology used to measure or estimate the amount of computation used for training.	☒	☒	□
Energy consumption (during training and inference)		**AIO**	**NCAs**	**DPs**
Amount of energy used for training:	Measured or estimated amount of energy used for training, reported in Megawatt-hours and recorded with at least two significant figures (e.g. $1.0x10^{2}$ MWh). If the amount of energy used for training cannot be estimated due to the lack of critical information from a compute or hardware provider, enter 'N/A'.	☒	☒	□
Measurement methodology:	In the absence of a delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies, describe the methodology used to measure or estimate the amount of energy used for training. Where the energy consumption of the model is unknown, the energy consumption may be estimated based on information about computational resources used. If the amount of energy used for training cannot be estimated due to a lack of critical information from a compute or hardware provider, the provider should disclose the type of information they lack. *[Recommended 100 words]*	☒	☒	□
Benchmarked amount of computation used for inference[1]:	Benchmarked amount of computation used for inference, reported in floating point operations, recorded with at least two significant figures (e.g. $5.1x10^{17}$ floating point operations).	☒	☒	□
Measurement methodology:	In the absence of a delegated act adopted in accordance with Article 53(5) AI Act to detail measurement and calculation methodologies, provide a description of a computational task (e.g. generating 100000 tokens) and the hardware (e.g. 64 Nvidia A100s) used to measure or estimate the amount of computation used for inference.	☒	☒	□

[1] This item relates to energy consumption during inference, which makes up the "energy consumption of the model" (Annex XI, 2(e), AI Act) together with energy consumption during training. Since energy consumption during inference depends on more than just the model itself, the information required for this item is limited to relevant information depending only on the model, namely computational resources used for inference.

2. Copyright Code of Practice

a. Commitment to place general-purpose AI models on the Union market must put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed by rightsholders.
b. Commitment to proportionate measures should be commensurate and proportionate to the size of providers, taking due account of the interests of SMEs, including startups;
c. Commitment to to draw up and make publicly available sufficiently detailed summaries about the content used by the Signatories for the training of their general-purpose AI models, according to a template to be provided by the AI Office

This Copyright code of Conduct is open to even greater challenges than the Transparency section, but this is because of the complexity of copyright law as the EU had little choice but to observe that the law varies throughout the world and that there was limited scope as a result for the requirement for providers to commit to copyright law.

Providers that place general-purpose AI models on the Union market must put in place a policy to comply with Union law on copyright and related rights, and in particular to identify and comply with, including through state-of-the-art technologies, a reservation of rights expressed by rightsholders pursuant to Article 4(3) of Directive (EU) 2019/790. It also does not affect agreements between the Signatories and rightsholders authorising the use of works and other protected subject matter

The Code is express that it does not affect the application and enforcement of Union law on copyright and related right which is for the Member States courts and the Court of Justice of the European Union to interpret, having regard in particular to Directive 2001/29/EC, Directive (EU) 2019/790 and Directive 2004/48/EC, and providing for exclusive rights that are preventive in nature and thus is based on prior consent save where an exception or limitation applies whilst at the same time providing an exception or limitation for text and data mining (TDM)[10] which is carried out via lawful access and not in breach of any expressly reservation of right[11] made in an appropriate manner[12].

As a result, the Copyright Code requires GPAI providers simply to implement measures to ensure compliance with EU copyright and related rights, although inevitably because of the jurisdictional nature of copyright, this will remain a grey area. For example, where a GPAI is trained in the USA according to US copyright law and caselaw, then the use of the GPAI in Europe will not infringe, in the majority of usages, EU copyright and related rights unless express provisions to protect that material apply..

Where a model is trained in the US but is fine-tuned in the EU, then the party carrying out the fine tuning will need to comply with EU copyright and related rights in relation to the training data used for fine-tuning.

[10] Article 4(1) of Directive (EU) 2019/790
[11] pursuant to Article 4(3) of Directive (EU) 2019/790
[12] i.e. in a conventional and recognised machine readable format

Most importantly, the Code is explicit in having no effect on the application and enforcement of European Union law in relation to copyright and related rights and notes that it is for the Courts of the Member States and ultimately the Court of Justice of the European Union to determine and interpret rights and obligations under copyright and related rights.

The Code allows providers "reproduce and extract only lawfully accessible copyright-protected content when crawling the World Wide Web" and second, that they "identify and comply with rights reservations when crawling the World Wide Web" and this reflects the recent LAION court case where material not behind passwords and paywalls was determined to be fair game for training data[13].

1. **Commitment to observe Copyright[14]**
   Providers must be prepared, for GPAI (general-purpose AI) models they place on the Union market, to:
   a) comply with Union law on copyright and related rights; and
   b) ensure that state-of-the-art technologies providing a reservation of rights are observed[15]

1.1 Commitment to create a Copyright Policy
Providers must draw up, keep up-to-date, and implement an internal policy for GPAI (general-purpose AI) models they place on the Union market to:
a) comply with Union law on copyright and related rights; and
b) ensure that state-of-the-art technologies providing a reservation of rights are observed[16]

Providers must also assign responsibilities within their organisation for the implementation and overseeing of this policy.

Providers are also encouraged (but not required) to make publicly available and keep up-to-date a summary of their copyright policy.

[13] Germany - Hamburg District Court, 310 O.22723, LAION v Robert Kneschke, [27 September 24]. LAION, a non-profit organisation, did not infringe copyright law by creating a dataset for training artificial intelligence (AI) models through web scraping publicly available images, as this activity constitutes a legitimate form of text and data mining (TDM) for scientific research purposes. The photographer Robert Kneschke unsuccessfully sued LAION for infringed his copyright by reproducing one of his images without permission as part of the training dataset creation process. LAION created this dataset by aggregating publicly available images and their corresponding textual descriptions. The dataset was made publicly available for free and could be used to train AI models. It was not disputed that LAION had downloaded a copy of Kneschke's image that was available in low resolution and watermarked. The Court of Hamburg dismissed the lawsuit ruling out a copyright infringement as LAION's reproduction of Kneschke's image was covered by the TDM exception for scientific research under Article 3 of the Digital Single Market (DSM) Directive, implemented in German law by Section 60d of the Act on Copyright and Related Rights (Urheberrechtsgesetz 'UrhG'). (Section 60d applying as the TDM was for scientific research purposes as LAION was a non-profit organisation and the dataset was created for non-commercial purposes and published free of charge, i.e. in the public interest.) It is likely that even if not NFP, then s44 may have applied. The Court also found that web-scraping protected works does not have commercial purposes even if it is ultimately aimed at generating (through AI) identical or similar products that will compete with them. It should be noted that the general exception for text and data mining – unlike the more specific exception for text and data mining for the purposes of scientific research – permits the rights holder to reserve the use of its work for text and data mining through an express declaration,so, obiter dictum (non-binding opinion), suggested that a reservation of rights could be made in natural language if machine-readable, noting that technologies are able to detect opt-outs expressed in natural language since 2021, meaning that robot.txt exclusions would also be valid; however the problem for artists is that any such language would exclude inclusion in google searches too.
https://www.wipo.int/wipolex/en/text/592042. This case broadly follows copyright law findings in the US where recognition of robot.txt files and notrain.txt files is followed but otherwise if you put your material on the web, it is for training, whether by humans or AI.

[14] Article 53(1)(c) AI Act

[15] pursuant to Article 4(3) of Directive (EU) 2019/790 which states in relation to Text and Data Mining (TDMs) "The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online".

[16] pursuant to Article 4(3) of Directive (EU) 2019/790 which states in relation to Text and Data Mining (TDMs) "The exception or limitation provided for in paragraph 1 shall apply on condition that the use of works and other subject matter referred to in that paragraph has not been expressly reserved by their rightholders in an appropriate manner, such as machine-readable means in the case of content made publicly available online".

This is a notable change from previous drafts where it was a stronger requirement and has been watered down as a result of lobbying by AI developers and providers who are now simply "*encouraged*" to make these public[17].

1.2 Reproduce and Extract only lawfully accessible copyright protected content when crawling the World Wide Web.

Providers must ensure that if they use web-crawlers or have such web-crawlers used on their behalf to scrape or otherwise compile data for the purpose of text and data mining as defined in Article 2(2) of Directive (EU) 2019/790 and the training of their general-purpose AI models:
(i) they only reproduce and extract lawfully accessible works and other protected subject matter:
(ii) not to circumvent effective technological measures as defined in Article 6(3) of Directive 2001/29/EC that are designed to prevent or restrict unauthorised acts in respect of works and other protected subject matter[18], and
(iii) exclude from their web-crawling websites that make available to the public content and which are, at the time of web-crawling, recognised as persistently and repeatedly infringing copyright and related rights.

1.3 Identify and Comply with rights when crawling the World Wide Web.

Providers must ensure that if they use crawlers then:
(i) these are able to understand state-of- the-art technologies, machine-readable reservations of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790; and
(ii) these can compile data for the purpose of text and data mining as defined in Article 2(2) of Directive (EU) 2019/790 and the training of their general-purpose AI models and comply with the exemptions under law[19];
(iii) these can read and follow instructions expressed in accordance with the Robot Exclusion Protocol (robots.txt), as specified in the Internet Engineering Task Force (IETF) Request for Comments No. 9309, and any subsequent version of this Protocol for which the IETF demonstrates that it is technically feasible and implementable by AI providers and content providers, including rightsholders,
(iv) these can identify and comply with other appropriate machine-readable protocols to express rights reservations pursuant to Article 4(3) of Directive (EU) 2019/790, for example through asset-based or location-based metadata, that have either have been adopted by international or European standardisation organisations, or are state-of-the-art, including technically implementable, and widely adopted by rightsholders, considering different cultural sectors, and generally agreed through an inclusive process based on bona fide discussions to be facilitated at EU level with the involvement of rightsholders, AI providers and other stakeholders[20].
v) they comply with the right of rightsholders to expressly reserve the use of works and other protected subject matter for the purposes of text and data mining pursuant to Article 4(3) of Directive (EU) 2019/790 in any appropriate manner, such as machine-readable means in the case of content made publicly available online or by other means whilst observing the exceptions for TDM.; and
vi) they to allow affected rightholders to obtain information about the web-crawlers employers, their robots.txt features and other measures adopted to observe machine readable rights reservations under Article 4(3) of Directive 2019/790 at the time of crawling. There must also be a means whereby rightholders are notifed when this information is updated.

In addition, where a provider also provides a search engine[21] then the new Code requires the search engine provider to take such measures to ensure that any rights reservation in relation to Text and Data Mining (TDM) does not directly lead to adverse indexing of the content, or the domain or URL in the search engine. (In other words, a search engine provider who also operates a GPAI system cannot penalize, whether by de-ranking or otherwise, the rightsholders or any site they have their works on for using a rights reservation or prohibition in relation to training or TDM.

[17] For the purpose of compliance with this measure, a dynamic list of hyperlinks to lists of these websites issued by the relevant bodies in the European Union and the European Economic Area will be made publicly available on an EU website.

[18] Including restriction of access imposed by subscription models or paywalls

[19] See Germany - Hamburg District Court, 310 O.22723, LAION v Robert Kneschke in above footnote

[20] Evidence to date is that any agreement between the various stakeholders is highly unlikely

[21] see Regulation 2022/2065 Article 3 (j)

When acquiring datasets from third parties (i.e., not directly web-crawled by the provider – for example when purchasing training data from 3rd parties), providers must make reasonable efforts to obtain adequate information about the protected content within those datasets. This includes information on whether the web crawlers used by the third parties to gather the data read and followed robots.txt instructions and complied with other relevant rights reservations. This aims to ensure that models trained on third-party data still uphold copyright compliance.

Previous provisions placing detailed obligations on providers in relation to non-web-crawled provisions have been removed following complaints by rightsholders that the provisions a) did not go far enough and b) were confusing and complaints by AI providers that the provisions were i) already covered by the existing requirements to observe copyright and related rights and ii) unworkably confusing.

1.4 Mitigating the risk of copyright infringing outputs

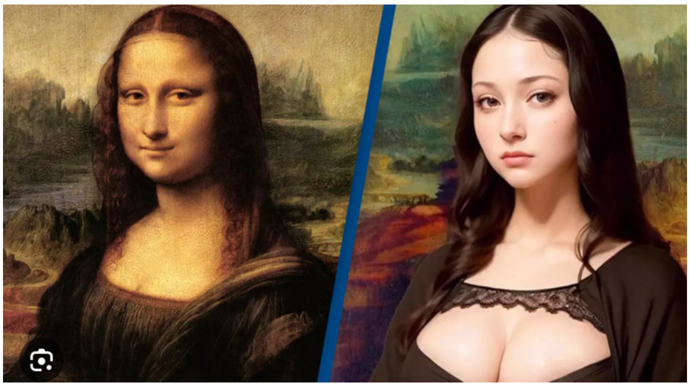
Providers are required to take steps to ensure that their systems, including any downstream systems,
a) cannot generate an output from its training data that infringes the rights in works or other subject matter protected by copyright law or related rights; and
b) implement safeguards that prevent the models generating outputs that reproduce the training content that is protected in copyright law and related rights in an infringing manner; and
c) prohibit use of their models in a copyright infringing manner via acceptable use policies, terms and conditions (and, for open source systems, also alert users to copyright infringing uses of the model and via documentation).

This reflects the fact that providers cannot use content acquired illegally, such as via pirate libraries and must also not circumvent effective technological protection measures (TPMs), such as paywalls or digital locks, to access content. They must also make reasonable efforts to avoid crawling from widely known piracy domains (e.g., websites primarily dedicated to hosting infringing content); however as the recent cases have determined, where artists have failed to use “no follow” and “no training” tags and have failed to provide their content behind paywalls, then the artists are deemed to have provided lawfully accessible content for training. (See also the measures highlighted above by Anthropic).

Under the code, providers must identify and comply with rights reservations (effectively any reasonable opt-out mechanisms) when crawling the World Wide Web, something that is linked to the Text and Data Mining (TDM) exception under Article 4(3) of the EU Copyright Directive (DSM Directive), which allows TDM unless a rightsholder has explicitly opted out via the Robot Exclusion Protocol, so providers must employ web-crawlers that read and follow instructions expressed in accordance with the Robots.txt file instructions. They must also make best efforts to identify and comply with other appropriate machine-readable protocols for expressing rights reservations (e.g., metadata tags, industry-standard protocols that emerge)[22].

The EU has resisted the lobbying of the rightsholders to restrict AI look-alikes whether these are fake news images or “artistic works in the style of X” and this reflects the fact that there is a public interest in the training of artists and that artists have for hundreds of years trained by learning to copy a work of a particular style and AI is no different. This also reflects the almost impossible task that the rightholders lobbying, if successful would have imposed on the copyright courts and judges to determine is a new work is sufficiently close to an artist’s style as to be infringing. This also reflects the few existing court cases at the date of the creation of the Code.

[22] This acknowledges the evolving landscape of opt-out mechanisms.



Providers must make reasonable efforts to mitigate the risk that the GPAI model generates copyright-infringing output. This includes risks that the model might "memorize" copyrighted training content to the extent that it repeatedly produces substantially similar or identical infringing outputs. This replaced a more problematic "overfitting" requirement in earlier drafts. This is a particularly difficult concept as it is an ill-defined area on copyright law. For many years, art students have been trained on the output of particular artists and encouraged to develop a style that emulates those particular artists, something expressly condoned by the artists. So it is not uncommon for artists to be able to paint in the style of Roy Lichtenstein or Andy Warhol or Francis Bacon or David Hockney and this is clearly not a copyright infringement except where a particular work is reproduced. So too, measure I.2.5 seeks to prevent AI from being able to regurgitate existing copyright works.

Providers and developers must also prohibit copyright-infringing uses in their acceptable use policies, terms and conditions, or equivalent documents for downstream users. This aims to prevent their models from being used for illegal activities, but there is nothing infringing in requesting a work "in the style of" a particular artist and the codes recognise this.

1.5 Points of Contacts and Complaints

There is also a requirement for providers to designate a point of e-communication for affected rightsholders as well as a requirement for a mechanism to be put into place so that rightsholders and their authorized representatives can submit "*sufficiently precise and adequately substantiated complaints"* concerning any noncompliance with the commitments of the copyright code.

Providers are required to act on "*sufficiently precise and adequately substantiated complaints"* in a diligent and non-arbitrary manner and within a reasonable time, so manifestly unfounded claims and claims which have previously bee responded to can be ignored by providers. (This does not however prevent any measures or remedies being sought as are set out under EU and national law law.

The "*sufficiently precise and adequately substantiated complaints"* is a measure which sets out a significantly higher bar than is seen in other fields such as domain name complaints where there is considerable abuse by representatives of rights holders who seek to simply notify infringement without any precise statement of how infringement arises and without any adequately substantiated evidence of the infringement.

3. Safety and Security Code of Practice

This is without doubt the largest by far of the Codes of Practice, running to approximately 40 pages, having been reduced down from 68 pages in an earlier draft. Each section therefore has objectives and commitments.

OBJECTIVES

The overarching objective of the Security Code of Practice is to:
a) improve the functioning of the internal market,
b) promote the uptake of human-centric and trustworthy artificial intelligence (“AI”),
c) ensure a high level of protection of health, safety, and fundamental rights
d) protect democracy and the rule of law against harmful effects of AI in the Union,
e) support AI innovation.

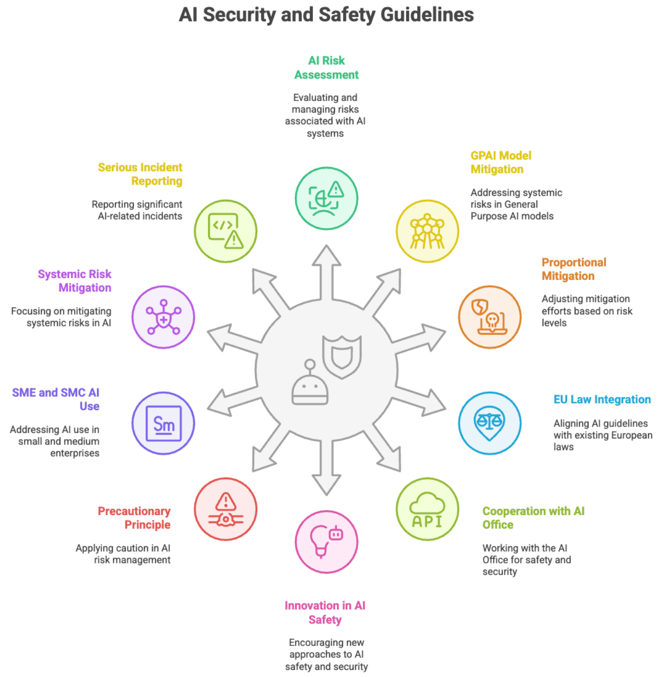
The Codes serve as a guiding document for demonstrating compliance with the obligations provided for in Articles 53 and 55 of the AI Act, while recognising that adherence to the Code does not constitute conclusive evidence of compliance with the obligations under the AI Act.

To ensure providers of general-purpose AI models comply with their obligations under the AI Act and to enable the AI Office to assess compliance of providers of general-purpose AI models who choose to rely on the Code to demonstrate compliance with their obligations under the AI Act.

Although draft v3 implied that compliance with the Code would be equivalent to compliance with the obligations under Articles 53 and 55 AI Act, the final version simply states that it is a guiding document for demonstrating compliance with Articles 53 and 55 AI Act but that adherence to the Code does not constitute conclusive evidence of compliance. It being intended that providers of GPAI models comply with their obligations under the AI Act. The Code is also intended to enable the AI Office to assess compliance.

Whilst the Code of Practice went from 16 commitments to 10 commitments[23], much of this was a result of consolidation of principles and removal of overlap although the final version removes only one major commitment which is arguably incorporated as a result of the more demanding wording, as well as relocation of technical content to annexes.

[23] Effectively 10 commandments, although no biblical connection is intended.


AI Security and Safety Guidelines
AI Risk Assessment
Evaluating and managing risks associated with AI systems
GPAI Model Mitigation
Addressing systemic risks in General Purpose AI models
Proportional Mitigation
Adjusting mitigation efforts based on risk levels
EU Law Integration
Aligning AI guidelines with existing European laws
Cooperation with AI Office
Working with the AI Office for safety and security
API
Innovation in AI Safety
Encouraging new approaches to AI safety and security
Precautionary Principle
Applying caution in AI risk management
SME and SMC AI Use
Addressing AI use in small and medium enterprises
Sm
Systemic Risk Mitigation
Focusing on mitigating systemic risks in AI
Serious Incident Reporting
Reporting significant AI-related incidents

The Security and Safety Guidelines are built around the following principles:

(a) AI Risk being assessed with Appropriate Lifecycle Management.
(b) Risk Assessment and Mitigation covering GPAI Models with systemic risk and not all AI systems
(c) Mitigation proportional to Systemic Risks.
(d) Integration with Existing EU Laws.
(e) Principle of Cooperation with AI Office
(f) Principle of Innovation in AI Safety and Security.
(g) Precautionary Principle for AI Risk.
(h) AI use by Small and medium enterprises (“SMEs”) and small mid-cap enterprises (“SMCs”).
(i) Interpretation with the objective to assess and mitigate systemic risks.
(j) Serious Incident Reporting.

1 Principle of Appropriate Lifecycle Management.

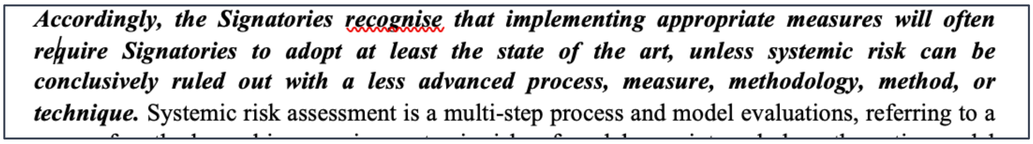
Providers GPAI models with systemic risk (GPAISRs) are required to :
a) continuously assess and mitigate systemic risks along the entire model lifecycle;
(This includes any period of development that occurs before a model has been placed on the market as well as after placement on the market).
b) co-operate with and take into account all relevant persons impacted along the AI value chain;
c) ensure that systemic risk management is made future-proof by regular updates in the model capabilities and risk profiles[24].

Although earlier versions allowed Providers to determine “appropriate steps” depending on their assessment of risk, leaving that interpretation largely in the hands of the Providers, in the final version of this code it was stated that implementing appropriate measures will usually

[24] (see recitals 114 and 115 AI Act)

require Providers *to adopt **at least** the state of the art* unless systemic risk can be **conclusively** ruled out with a less advanced process, measure, methodology, method, or technique.

This implementation of "at least the State of the Art" for assessing risk is a very high obligation and will mean that Providers will need to consider what is the State of the Art for assessing risk and must meet that level, if not exceed it. It also means that what meets the requirement at a given date is unlikely to meet the requirement some 6 months later as the State of the Art in risk assessment will have moved on.



Accordingly, the Signatories recognise that implementing appropriate measures will often require Signatories to adopt at least the state of the art, unless systemic risk can be conclusively ruled out with a less advanced process, measure, methodology, method, or technique. Systemic risk assessment is a multi-step process and model evaluations, referring to a

This is particularly so when coupled with the exception being limited to conclusively ruled out and this means that unless the Provider is able to guarantee that the systemic risk is so low that something less than latest State of Art is adequate to rule out risk, then the Provider will not meet the requirement. This is particularly so given the use of the phrase "..*at least the State of the Art*" as this phrase implies that the Providers are required no only to meet the State of the Art tests but to go beyond them. (Arguably, once on provider goes beyond the State of the Art test, then arguably that becomes the latest standard to meet the State of the art, an ever increasingly difficult standard to meet and potentially requiring the Provider to be aware of all of the current standards and to meet a standard of one of them and incorporate other elements from other standards. (Whether the drafting committee intended this standard is open to speculation).

This Principle also notes that

> *"Systemic risk assessment is also stated to be "a multi-step process and model evaluations, referring to a range of methods used in assessing systemic risks of models, are integral along the entire model lifecycle".*

When systemic risk mitigations are implemented, it is also expected that the Providers will recognise the importance of continuously assessing their effectiveness.

2 Principle of Contextual Risk Assessment and Mitigation.

The Safety and Security Chapter is only relevant for providers of general-purpose AI models (GPAIs) with systemic risk and not AI systems, but requires that the assessment and mitigation of systemic risks should include, as reasonably foreseeable, the system architecture, other software into which the model may be integrated, and the computing resources available at inference time because of their importance to the model's effects, for example by affecting the effectiveness of safety and security mitigations.

This effectively means that the Provider of a GPAI that has systemic risk (GPAISRs) will be required to consider the operational nature of their clients who may be introducing that AI into their operations by combination with other software, which will raise important know your client obligations on the provider as well as business secrecy practice and confidentiality protection on both parties, again something that potentially places a very high burden on Providers.

3 Principle of Proportionality to Systemic Risks.

The Code recognises that the assessment and *mitigation* of systemic risks should be proportionate to the risks involved with the relevant AI model (Article 56(2), point (d) AI Act). It also recognizes that the degree of scrutiny in systemic risk assessment and mitigation should be proportionate to the systemic risks at the relevant points along the entire model lifecycle, and that the level of detail in documentation and reporting should reflect this, meaning that the higher the assessed systemic risk, the higher in the document burden and the more detail must be included, but the Code also recognises that while systemic risk assessment and mitigation is iterative and continuous, it does not need to duplicate assessments that are still appropriate to the systemic risks stemming from the model, thus avoiding pointless duplication of assessments.

4 **Principle of Integration with existing EU Laws.**
Known as the harmonization and integration principle, the Code recognizes that where other EU laws provide international standards that cover the provisions of this Code then they can be relied on to avoid duplication, Arguably, however, where there is overlap between Latest State of Art and existing standards both standards should be met.

5 **Principle of Co-operation.**
The Code authors recognised that assessment and mitigation of systemic risk is something that requires significant investment of time and resources for a Provider and allows Providers to take advantages of collaborative efficiency, such as the sharing of model evaluations methods and/or infrastructure and this will extend to such co-operation with licensees, downstream modifiers, and downstream providers as well as engagement with experts and other representatives in the corporate and academic fields as well as other relevant stakeholders.

This, it is recognized may also require agreements to share information relevant to systemic risk assessment and mitigation, while ensuring proportionate protection of sensitive information and compliance with applicable Union law, so that commercial confidentiality can be maintained even in the event of a systemic risk being identified; however the Code also reflects the importance of cooperating with the AI Office (Article 53(3) AI Act) to foster collaboration between providers of general-purpose AI models with systemic risk, researchers, and regulatory bodies to address emerging challenges and opportunities in the AI landscape.

This however will create, for many Providers and their external experts and other representatives and consultees, an inherent conflict between the contractual obligation of secrecy of Provider confidentiality in relation to security and safety and their duty as corporate representatives and officers to the Provider company, namely that systemic concerns, problems and risks are kept secret and confidential to the Provider and the intention of the Code that systemic concerns, problems and risks that are identified are notified to the AI Office.

The Code and EU AI Act seems to imply that part of the function of the AI Office is a sharing and dissemination of a pool of critical knowledge, to assist with identifying and mitigating systemic risk, in order to foster collaboration, something that in the majority of cases is likely to be highly controversial even if the information is subjectively anonymized. (It being likely that any meaningful disclosure would be such that the disclosing provider could be identified with the consequential damage to share price and commercial opportunity). It will therefore be interesting to see if any meaningful disclosures to the AI office under this Code will be made, despite the risk of a finding of non-compliance by a non-disclosure.

6 **Principle of Innovation in AI Safety and Security.**
The authors of the Code recognised that determining the most effective methods for understanding and ensuring the safety and security of general- purpose AI models with systemic risk remains an evolving challenge and sought to provide a mechanism by which providers of GPAISRs are encouraged to advance the state of the art in AI safety and security and related

processes and measures.

They also recognised that advancing the state of the art also includes developing targeted methods that specifically address risks while maintaining beneficial capabilities (such as, for example, mitigating biosecurity risks without unduly reducing beneficial biomedical capabilities, or facilitating pharmaceutical advances whilst avoiding creation of new unknown poisons).

As such, where this systemic risk is identified then greater technical effort and innovation is required to maximise the innovation and there is an expectation that providers of GPAISRs targeted methods will demonstrate equal or superior safety or security outcomes and that these may be achieved via alternative means that achieve greater efficiency, and that where such innovations are identified they should be recognised as advancing the state of the art in AI safety and security and meriting consideration for wider adoption. In the majority of cases however, the disclosure of the alternative means for achieving the goal with greater efficiency would have a material commercial value and no mechanism has been developed for compensation to the Provider for any such disclosure, and therefore it is unlikely that any such advance would be readily disclosed (due to the legal obligations to the provider company and its shareholders), at least prior to the technique becoming public knowledge.

7 **Precautionary Principle.**

The Code also recognises the important role of the Precautionary Principle, particularly for systemic risks for which the lack or quality of scientific data does not yet permit a complete assessment, and envisages that extrapolation of current adoption rates and research and development trajectories of models should be taken into account for the identification of systemic risks. This does however allow a degree of gaming of the Codes to reduce the efficacy of assessment of systemic risks for commercial advantage.

The history of Precautionary Principles in legislative measures has rarely been adopted without a degree of compulsion via significant fines for non-compliance.

8 **Adjustment for Small and medium enterprises (“SMEs”) and small mid-cap enterprises (“SMCs”).**

In an attempt to account for differences between size and capacity of providers of GPAISRs, the Code aims at a simplified way of proportionate compliance for startups, SMEs and SMCs, including startups, as is typical with EU Legislation. The idea is that SMEs and SMCs may be exempted from some reporting commitments.

SMEs and SMCs are typically more innovative in the developmental fields and may present as large a risk to safety and security in AI areas as large enterprises due to that innovation, although the nature of that innovation is also often desirable and reflects the desire of the Code to stimulate innovation. The feat that SMEs and SMCs would be stifled by having to comply with the stringent requirements of the Code of Practice is very well founded as the evidence to date is that, whilst they recognize the need for safety and security and have fledgling red teams and other security apparatus, they do not have the resources that are available in their large AI competitors and often rely upon a small team of specialists.The difficulty for the Code is that startups can grow rapidly from SMEs to very large capitalized entities by a single round of funding but may not have the capacity for some time after that funding to be able to meet the same commitments as large enterprises and this will continue to pose a material risk to safety and security for AI.

9 **Interpretation.**

All of the commitments and measures under the Codes are to be interpreted in light of the objective to assess and mitigate systemic risks, recognizing that due to the rapid pace of AI development, a purposive interpretation[25] must be applied to the focus on systemic risk assessment and mitigation to future proof the legislation and interpretation must also reflect good faith in light of:

(1) the probability and severity of harm pursuant to the definition of 'risk' in Article 3(2) AI Act; and
(2) the definition of 'systemic risk' in Article 3(65) AI Act.

Interpretational guidelines will also be issued by the AI Offices.

10 Serious Incident Reporting.

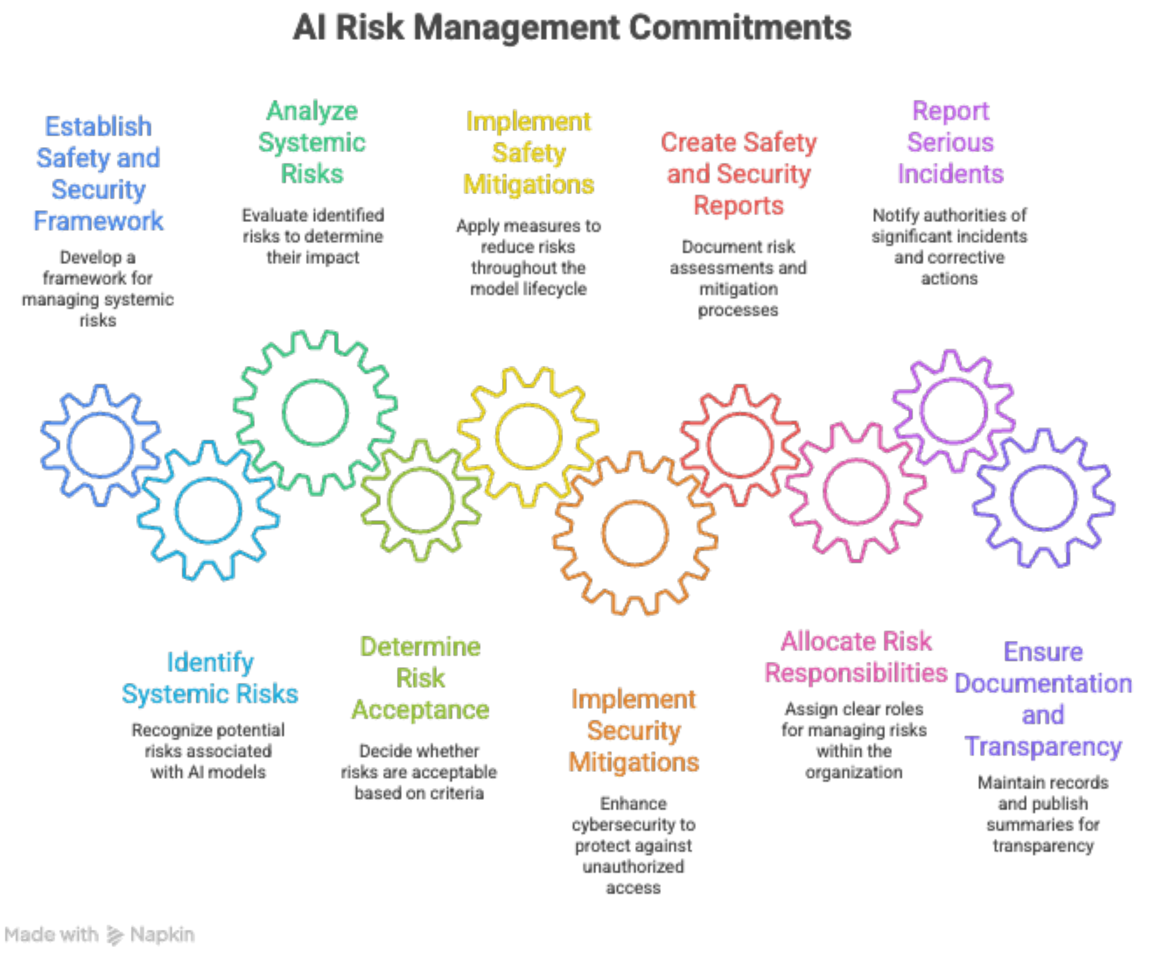
The Code provides that the reporting of a serious incident is not an admission of wrongdoing and recognises that relevant information about serious incidents will usually be documented, and reported at the model level in retrospect; however it seeks to proactively track, document and report in real time, although this is highly likely to be possible on grounds of corporate obligations, the rules of financial markets and the need for internal secrecy whilst a full investigation and guardrails are adjusted and safeguarding measures are put into place as well as requirements of potential litigation management, and in this respect the Code intention is exceedingly naaive.

The Code has concerns that, after a serious incident has occurred, critical information that could directly or indirectly be lost, overwritten, obscured, deliberately deleted or fragmented during investigation and the Code seeks to impose processes and measures to keep track of and document relevant information.

Even more naiively the Code seeks to impose processes and measures to keep track of and document relevant information before serious incidents occur. If a provider of GPAISRs was aware of a pending serious incident, then it would be entirely justified in taking steps to avoid the serious incident, in which case the serious incident does not arise and therefore there is no serious incident to report.

[25] The purposive approach to interpreting legislation is common in EU law and looks beyond the strict interpretation approach that looks at the words of the legislation at the purpose behind it, and the legislation is seen as a skeleton of the law for the judges to flesh out in time. The purposive approach has its roots in legal systems which are based on civil codes and is sometimes referred to as the teleological approach. It is used in EU law.

THE TEN COMMITMENTS


AI Risk Management Commitments
Establish Safety and Security Framework
Develop a framework for managing systemic risks
Analyze Systemic Risks
Evaluate identified risks to determine their impact
Implement Safety Mitigations
Apply measures to reduce risks throughout the model lifecycle
Create Safety and Security Reports
Document risk assessments and mitigation processes
Report Serious Incidents
Notify authorities of significant incidents and corrective actions
Identify Systemic Risks
Recognize potential risks associated with AI models
Determine Risk Acceptance
Decide whether risks are acceptable based on criteria
Implement Security Mitigations
Enhance cybersecurity to protect against unauthorized access
Allocate Risk Responsibilities
Assign clear roles for managing risks within the organization
Ensure Documentation and Transparency
Maintain records and publish summaries for transparency
Made with Napkin

There are Ten Commitments that make up the Safety and Security Code of Practice:

Commitment 1 A Safety and Security Framework
Designed to identify and outline the systemic risk management processes and the measures that are taken to ensure the systemic risks stemming from their models are controlled and risks minimized to an acceptable level[26].

Commitment 2 Systemic risk identification
A commitment to identifying the systemic risks stemming from the model, including facilitating systemic risk analysis (pursuant to Commitment 3) and systemic risk acceptance determination (pursuant to Commitment 4)[27].

Commitment 3 Systemic risk analysis
A commitment to analysing each identified systemic risk (pursuant to Commitment 2), with the purpose of facilitating systemic risk acceptance determination (ref: Commitment 4)[28].

[26] Articles 55(1) and 56(5), and recitals 110, 114, and 115 AI Act
[27] Article 55(1) and recital 110 AI Act
[28] Article 55(1) and recital 114 AI Act

Commitment 4 Systemic risk acceptance determination

A commitment to specifying systemic risk acceptance criteria and determining whether the systemic risks stemming from the model are acceptable (as specified in Measure 4.1) with providers committing to decide whether or not to proceed with the development, the making available on the market, and/or the use of the model based on the systemic risk acceptance determination (as specified in Measure 4.2)[29].

Commitment 5 Safety mitigations

A commitment to implement appropriate safety mitigations along the entire model lifecycle, as specified in the Measure for this Commitment, to ensure the systemic risks stemming from the model are acceptable (pursuant to Commitment 4)[30].

Commitment 6 Security mitigations

A commitment to implement an adequate level of cybersecurity protection for all models and their physical infrastructure along the entire model lifecycle to ensure the systemic risks stemming from their models that could arise from unauthorised releases, unauthorised access, &/or model theft are acceptable (pursuant to Commitment 4)[31].

Commitment 7 Safety and Security Model Reports

A commitment to reporting to the AI Office information about their model and their systemic risk assessment and mitigation processes and measures by creating a Safety and Security Model Report ("Model Report") before placing a model on the market (as specified in Measures 7.1 to 7.5). Further, Signatories commit to keeping the Model Report up-to-date (as specified in Measure 7.6) and notifying the AI Office of their Model Report (as specified in Measure 7.7)[32].

Commitment 8 Systemic risk responsibility allocation

A commitment to: (1) defining clear responsibilities for managing the systemic risks stemming from their models across all levels of the organisation (as specified in Measure 8.1); (2) allocating appropriate resources to actors who have been assigned responsibilities for managing systemic risk (as specified in Measure 8.2); and (3) promoting a healthy risk culture (as specified in Measure 8.3)[33].

Commitment 9 Serious incident reporting

A commitment 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[34].

Commitment 10 Additional documentation and transparency

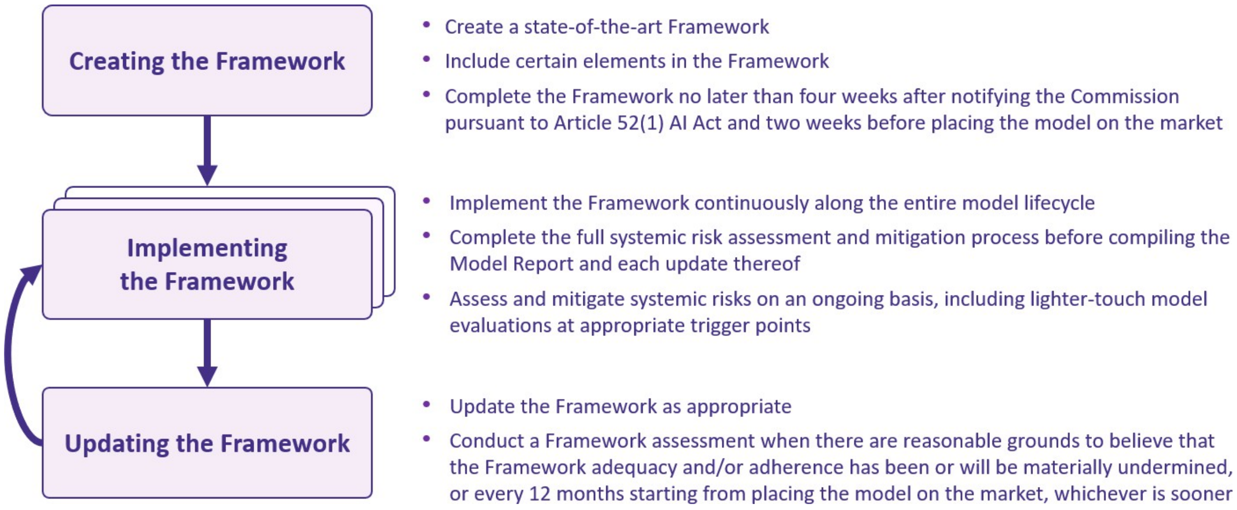
A Commitment to documenting the implementation of this Chapter (as specified in Measure 10.1) and publish summarised versions of their Framework and Model Reports as necessary (as specified in Measure 10.2)[35].

[29] Article 55(1) AI Act.
[30] Article 55(1) and recital 114 AI Act
[31] Article 55(1), and recitals 114 and 115 AI Act
[32] Articles 55(1) and 56(5) AI Act
[33] Article 55(1) and recital 114 AI Act
[34] Article 55(1), and recitals 114 and 115 AI Act
[35] Articles 53(1)(a) and 55(1) AI Act

COMMITMENT 1
A SAFETY AND SECURITY FRAMEWORK[36]

To identify and outline the systemic risk management processes and the control measures taken in relation to systemic risks and to minimize these to an acceptable level.

A three step adoption process for the Safety and Security Framework:



- Create a state-of-the-art Framework
- Include certain elements in the Framework
- Complete the Framework no later than four weeks after notifying the Commission pursuant to Article 52(1) AI Act and two weeks before placing the model on the market

- Implement the Framework continuously along the entire model lifecycle
- Complete the full systemic risk assessment and mitigation process before compiling the Model Report and each update thereof
- Assess and mitigate systemic risks on an ongoing basis, including lighter-touch model evaluations at appropriate trigger points

- Update the Framework as appropriate
- Conduct a Framework assessment when there are reasonable grounds to believe that the Framework adequacy and/or adherence has been or will be materially undermined, or every 12 months starting from placing the model on the market, whichever is sooner

Figure 1. Process for creating, implementing, and updating Frameworks.
The text of the Commitments and Measures takes precedence. (Source EC).

Note: There is also a commitment to notify the AI Office of the Framework.

1.1 CREATING THE FRAMEWORK

Creating the Framework must take into account the models being developed, made available on the market, and/or used and it must contain a high-level description of implemented and planned processes and measures for systemic risk assessment and mitigation. The Framework must contain[37]:

(1) a description and justification of the trigger points (and their usage) when the conducting of additional lighter-touch model evaluations will occur – for the entire model lifecycle[38].

(2) A determination of whether systemic risk is acceptable[39] including:

   (a) a description of the systemic risk acceptance criteria, including justification of criteria and any systemic risk tiers, and their usage;

   (b) a high-level description of what safety and security mitigations are in place for each systemic risk tier reached;

   (c) for each systemic risk tiers specified, estimates of timelines when it is reasonably foreseen that the relevant model will exceeds the highest systemic risk tier already reached by any existing models[40]; and

[36] Articles 55(1) and 56(5), and recitals 110, 114, and 115

[37] Measure 1.1

[38] See also implementation measures.

[39] Referencing Commitment 4

[40] These estimates may consist of time ranges or probability distributions; and may take into account aggregate forecasts, surveys, and other estimates produced with other providers and must be supported by justification and statements of any underlying assumptions and uncertainties

(d) a description of any influences from external actors[41] which have influenced development of the system, the making available on the market, and/or use of models. (for these purposes bona-fide independent external evaluations are excluded);

(3) a description of how systemic risk responsibility and mitigation is allocated[42]; and

(4) a description of the process by which the Framework is updated and any confirmation steps.

TIMESCALE: The Framework must be notified no later than four weeks after an Article 52(1) AI Act notification and no later than two weeks before placing the model on the market.

1.2 IMPLEMENTING THE FRAMEWORK

It is necessary, along the entire model lifecycle, to continuously:

(1) assess the systemic risks stemming from the model by:

(a) conducting lighter-touch model evaluations[43] at appropriate trigger points defined in terms of, e.g. time, training compute, development stages, user access, inference compute, and/or affordances;

(b) conducting post-market monitoring after placing the model on the market[44];

(c) taking into account relevant information about serious incidents[45]; and

(d) conducting a full systemic risk assessment and mitigation process[46]

(2) implement systemic risk mitigations taking into account the results of point (1), including addressing serious incidents as appropriate.

It is necessary to implement a full systemic risk assessment and mitigation process:

(1) identifying the systemic risks stemming from the model[47];

(2) analysing each identified systemic risk [48];

(3) determining whether the systemic risks stemming from the model are acceptable; and

(4) identifying, if the systemic risks are found not to be acceptable, the process to implemen safety and/or security mitigations[49], and re-assessing the systemic risks.

NOTE: A full systemic risk assessment and mitigation process MUST be carried out before placing the model on the market and whenever the conditions for model update reports arise.
There is a requirement to report their implemented measures and processes to the AI Office[50].

[41] This will include governments, lobby groups, academia etc

[42] Including having regard to Commitment 8

[43] these need not adhere to Appendix 3 (e.g. automated evaluations)

[44] see Post-Marketing Obligations and Measures

[45] see Commitment 9

[46] increasing the breadth and/or depth of assessment or conducting a full systemic risk assessment and mitigation process as is appropriate

[47] Ref Commitment 2

[48] Ref Commitment 2

[49] Commitments 5 and 6

[50] See Commitment 7

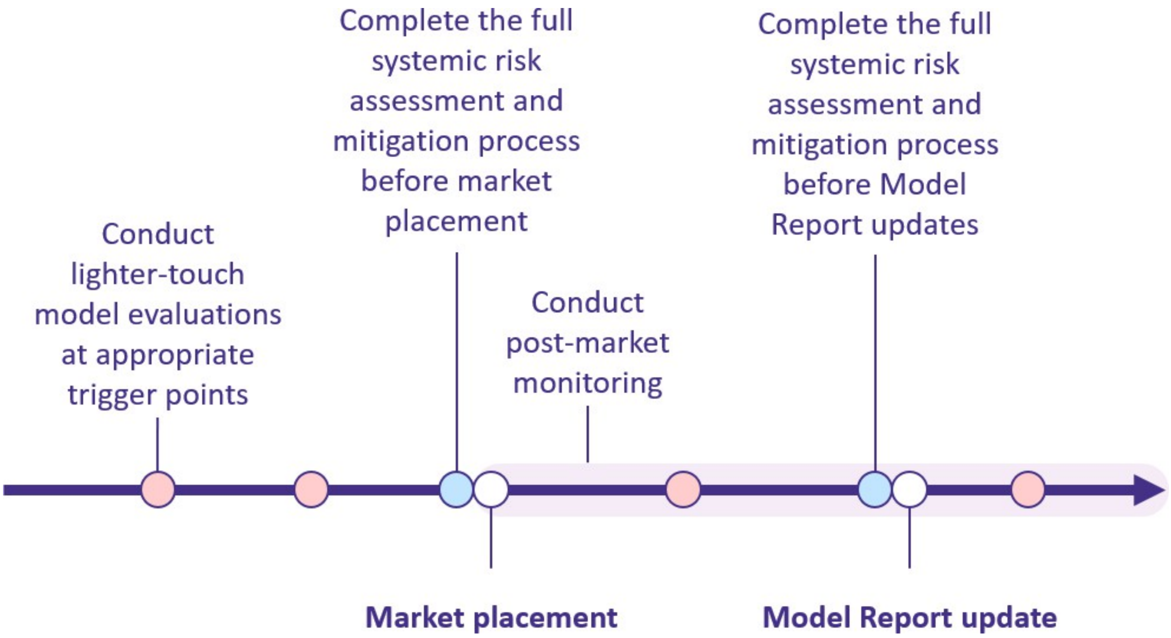

Conduct lighter-touch model evaluations at appropriate trigger points
Complete the full systemic risk assessment and mitigation process before market placement
Conduct post-market monitoring
Complete the full systemic risk assessment and mitigation process before Model Report updates
Market placement
Model Report update

Figure 2. Illustrative timeline of systemic risk assessment and mitigation along the model lifecycle. The text of the Commitments and Measures takes precedence.

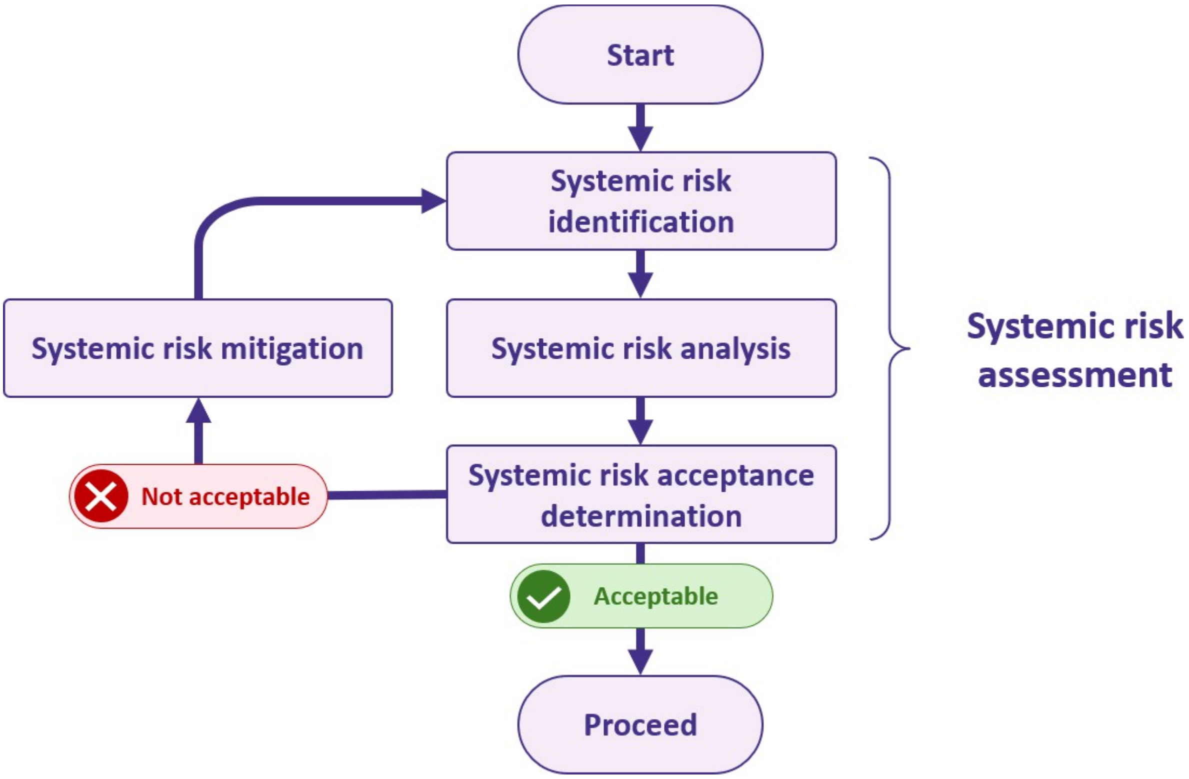


Figure 3. Full systemic risk assessment and mitigation process. The text of the Commitments and Measures takes precedence.

1.3 UPDATING THE FRAMEWORK

There is an obligation to update the Framework as appropriate, including without undue delay after a Framework assessment (see below) to ensure the information in the Framework is kept up-to- date and represents at least state-of-the-art, and where updates to the Framework occur then this must include a changelog[51] describing how and why the Framework has been updated including any responsibility changes.

It is necessary to conduct an appropriate Framework assessment, if there are reasonable grounds to believe that (i) the adequacy of the Framework has been or will be materially undermined, or
(ii) there has been a failure to adhere to the Framework and the steps therein.

The Framework must be updated every 12 months from the placing of the model on the market or any of the steps in (i) and (ii) above, whichever is sooner.

Examples:
1. If developing models change materially, and this reasonably foreseeably leads to the systemic risks not being acceptable.
2. If serious incidents and/or near misses occur involving their models or similar models, such that there is an indication that the systemic risks may have become unacceptable.
3. Systemic risks stemming from at least one model has changed or is likely to change materially[52], or at least one of model has developed or is likely to develop materially changed capabilities and/or propensities.

Framework adequacy
This requires an assessment of whether the processes and measures in the Framework are appropriate for the systemic risks stemming from the models and must take into account how the models are currently being developed, made available on the market, and/or used[53], and how they are expected to be developed.

Framework adherence
This is an assessment focused on the adherence to the Framework, including:
(a) any instances of, and reasons for, non-adherence to the Framework since the last Framework assessment; and
(b) any measures, including safety and security mitigations, that need to be implemented to ensure continued adherence to the Framework.
NB: If point(s) (a) and/or (b) give rise to risks of future non-adherence, it is necessary to make immediate remediation plans as part of the Framework assessment and adherence obligations.

Measure 1.4
Framework Notification
It is necessary to provide the AI Office with full access to their Framework, and updates thereof, within five business days of either being confirmed. Note that redacting will not be permitted.

[51] Including a version number and the date of change
[52] safety and/or security mitigations have become or are likely to become materially less effective
[53] Historically, currently or over the next 12 months

COMMITMENT 2
SYSTEMIC RISK IDENTIFICATION[54]

A commitment to identifying the systemic risks stemming from the model, including facilitating systemic risk analysis[55] and systemic risk acceptance determination[56]. Systemic risk identification involves two elements:

(1) following a structured process to identify the systemic risks stemming from the model (a System Risk Identification Process);; and

(2) developing systemic risk scenarios for each identified systemic risk.

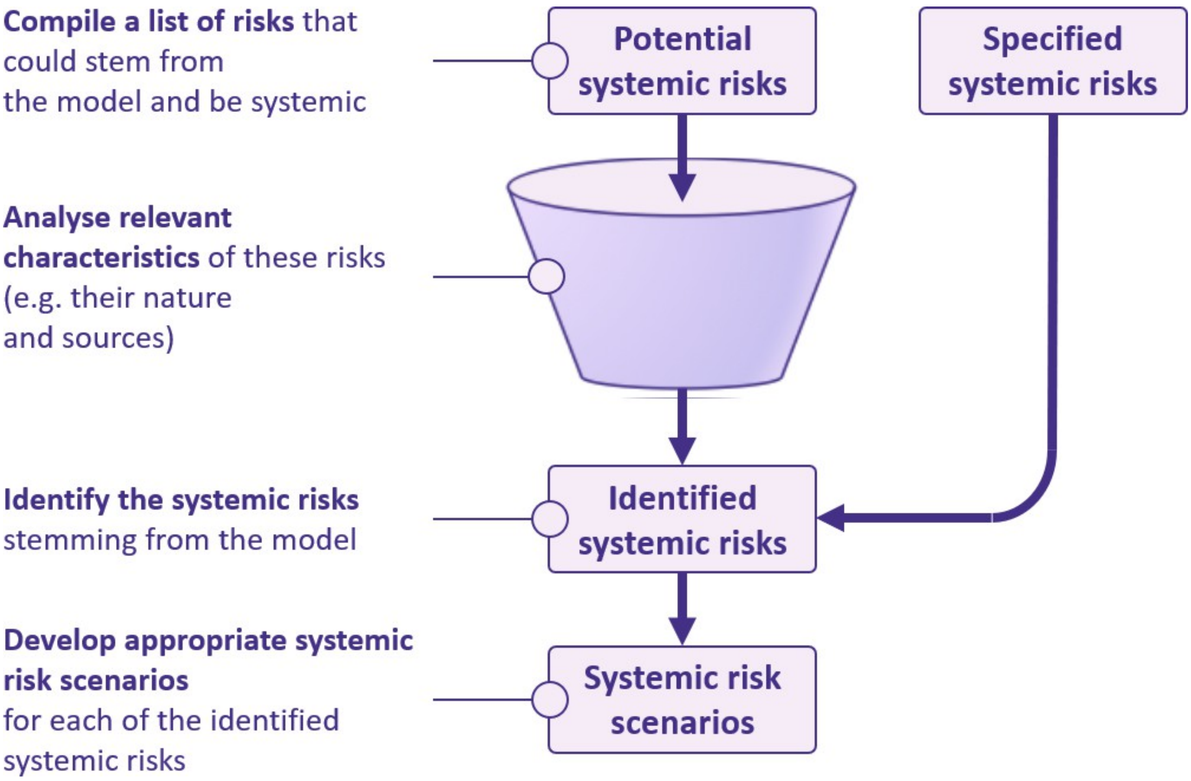
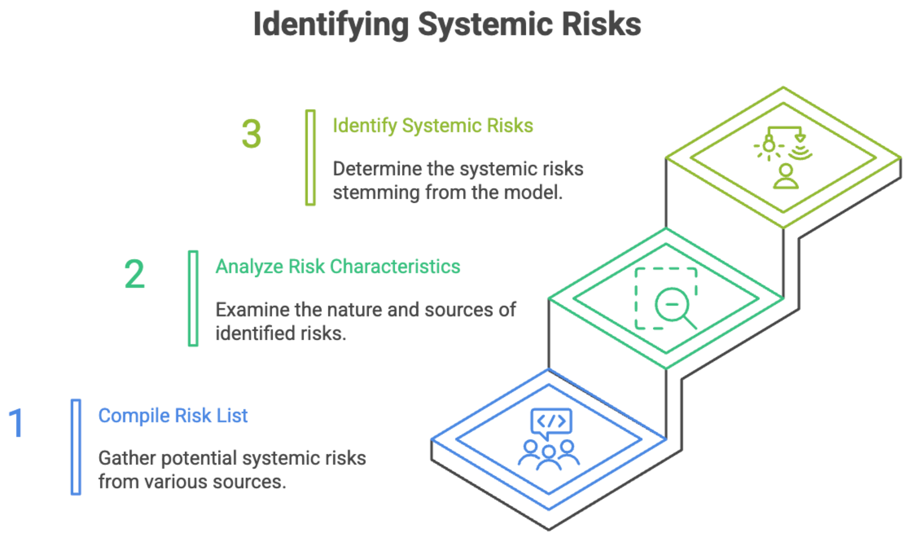

Compile a list of risks that could stem from the model and be systemic
Potential systemic risks
Specified systemic risks
Analyse relevant characteristics of these risks (e.g. their nature and sources)
Identify the systemic risks stemming from the model
Identified systemic risks
Develop appropriate systemic risk scenarios for each of the identified systemic risks
Systemic risk scenarios

Figure 4. Systemic risk identification process. The text of the Commitments and Measures takes precedence.

[54] Article 55(1) and recital 110 AI Act
[55] pursuant to Commitment 3
[56] pursuant to Commitment 4

Systemic Risk Identification Process


Identifying Systemic Risks
3
Identify Systemic Risks
Determine the systemic risks stemming from the model.
2
Analyze Risk Characteristics
Examine the nature and sources of identified risks.
1
Compile Risk List
Gather potential systemic risks from various sources.

The Systemic Risk Identification Process must identify:

(1) the systemic risks obtained, assessed and identified by:

(a) compiling a list of risks that could stem from the model and be systemic, based on the types of risks in Appendix 1.1, taking into account:

(i) model-independent information;

(ii) relevant information about the model and similar models, including information from post-market monitoring, and information about serious incidents and near misses[57]; and

(iii) any other relevant information communicated directly or via public releases by the AI Office, the Scientific Panel of Independent Experts, or other initiatives endorsed for this purpose by the AI Office[58];

(b) analysing relevant characteristics of the risks compiled pursuant to point (a), such as their nature (based on Appendix 1.2) and sources (based on Appendix 1.3); and

(c) identifying, based on point (b), the systemic risks stemming from the model; and

(2) the specified systemic risks in Appendix 1.4.

Measure 2.2 Systemic risk scenarios
It will also be necessary to develop appropriate systemic risk scenarios, including regarding the number and level of detail of these systemic risk scenarios, for each identified systemic risk (under the Systemic Risk Identification Process).

[57] pursuant to Commitment 9

[58] For example, , the International Network of AI Safety Institutes.

COMMITMENT 3
SYSTEMIC RISK ANALYSIS [59]

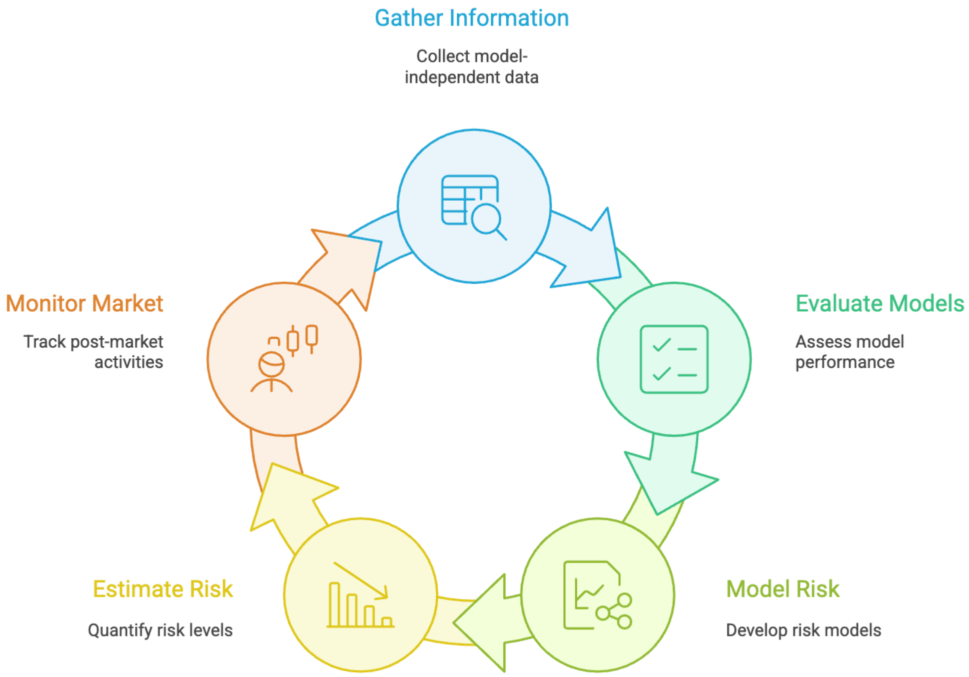
Following the **Systemic risk identification, this is a** commitment to analysing each identified systemic risk with the aim and purpose of facilitating Systemic Risk Acceptance Determination[60].

Systemic risk analysis involves 5 overlapping elements that may need to be implemented recursively:

(1) gathering model-independent information;
(2) conducting model evaluations;
(3) modelling the systemic risk;
(4) estimating the systemic risk; and
(5) a need to be conducting post-market monitoring.

The aim is to allow Providers to carry out analysis to determine the severity and probability of the systemic risks.

Systemic Risk Analysis Cycle

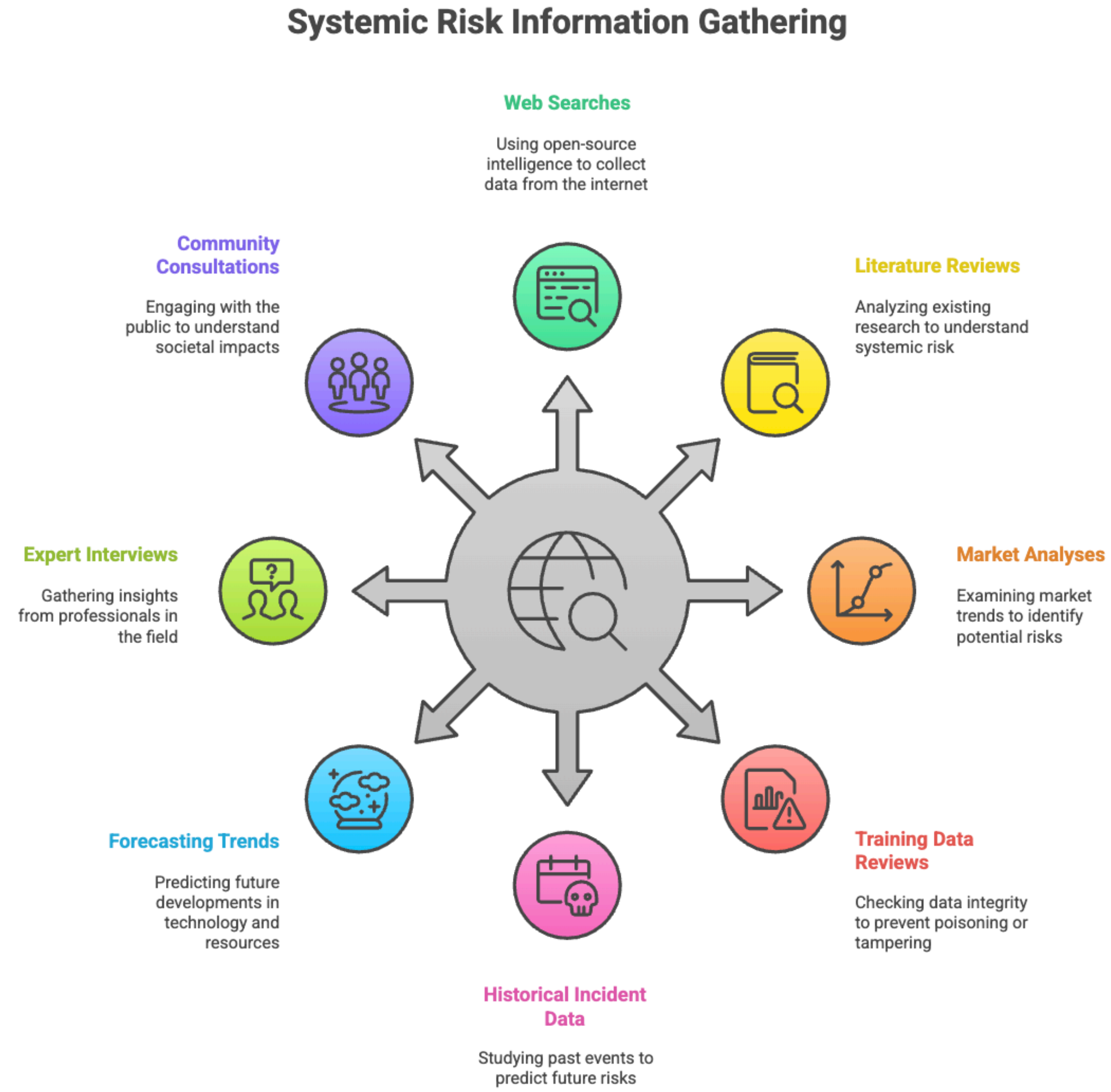


Measure 3.1 Model-independent information
This will search for and gather information as appropriate for the systemic risk, using methods such as:

(1) web searches[61] use of open-source intelligence methods in collecting and analysing information gathered from open sources;
(2) literature reviews;
(3) market analyses[62];
(4) reviews of training data[63] for example looking for indications of data poisoning or tampering;
(5) reviewing and analysing historical incident data and incident databases;

[59] Article 55(1) and recital 114 AI Act
[60] see Commitment 4
[61] use of open-source intelligence methods in collecting and analysing information gathered from open sources
[62] Typically focused on capabilities of other models available on the market
[63] for example looking for indications of data poisoning or tampering

(6) forecasting of general trends[64] For example, forecasts concerning the development of algorithmic efficiency, compute use, data availability, and energy use ;
(7) expert interviews and/or panels; and/or
(8) lay interviews, surveys, community consultations, or other participatory research methods investigating matters such as the effects of models on natural persons.

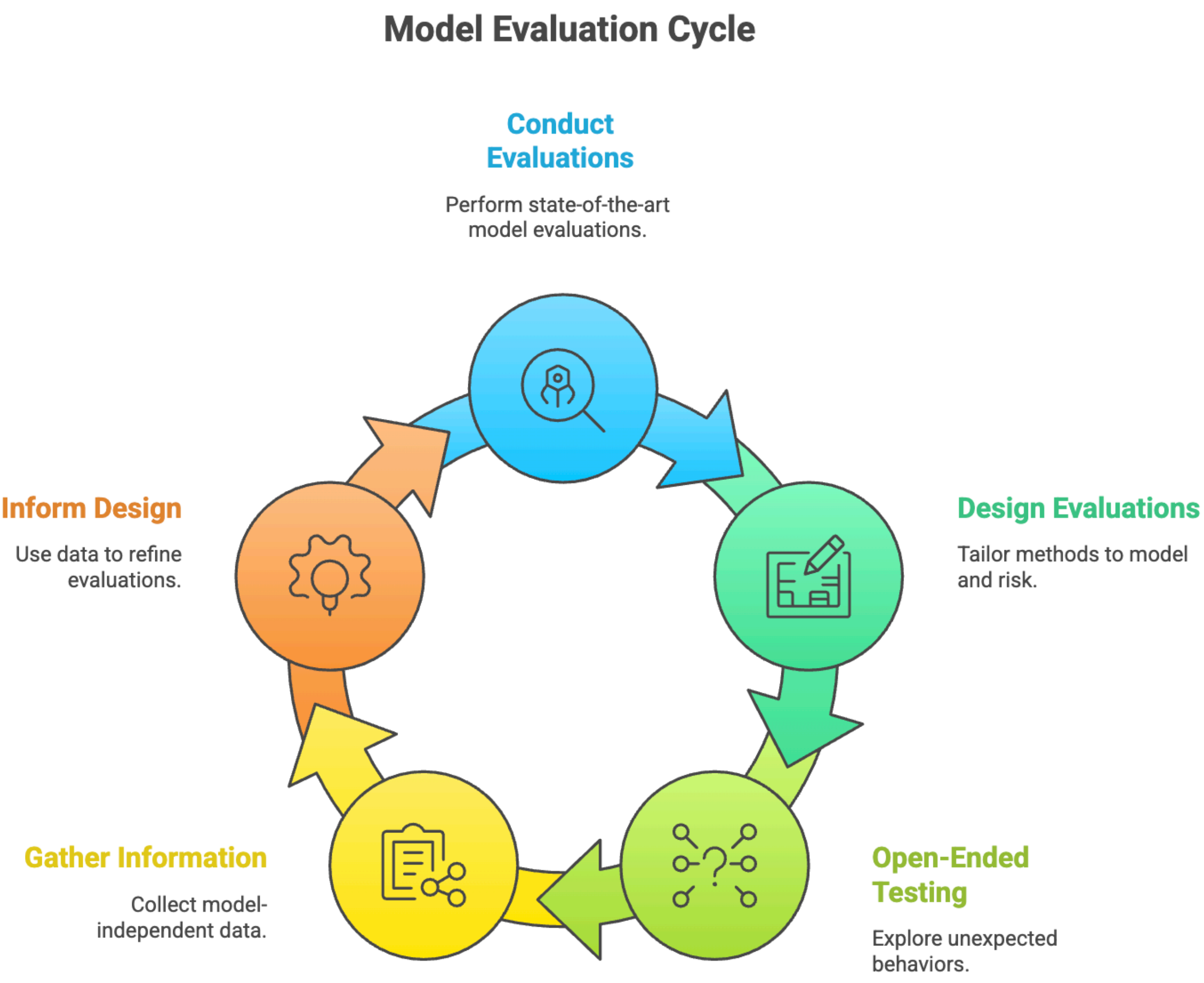

Systemic Risk Information Gathering
Web Searches
Using open-source intelligence to collect data from the internet
Literature Reviews
Analyzing existing research to understand systemic risk
Market Analyses
Examining market trends to identify potential risks
Training Data Reviews
Checking data integrity to prevent poisoning or tampering
Historical Incident Data
Studying past events to predict future risks
Forecasting Trends
Predicting future developments in technology and resources
Expert Interviews
Gathering insights from professionals in the field
Community Consultations
Engaging with the public to understand societal impacts

Measure 3.2 Model evaluations
It will be necessary to conduct **at least** state-of-the-art model evaluations in the modalities relevant to the systemic risk to assess the model's capabilities, propensities, affordances, and/or effects, as specified in Appendix 3.

This is designed to ensure that such model evaluations are designed and conducted using methods that are appropriate for the model and the systemic risk, and include open-ended testing of the model to improve the understanding of the systemic risk, with a view to identifying unexpected behaviours, capability boundaries, or emergent properties. Examples of model evaluation methods are: Q&A sets, task-based evaluations, benchmarks, red-teaming and other methods of adversarial testing, human uplift studies, model organisms, simulations, and/or proxy evaluations for classified materials. Further, the design of the model evaluations

[64] For example, forecasts concerning the development of algorithmic efficiency, compute use, data availability, and energy use

will be informed by the model-independent information gathered as part of the model independent information gathering process..



Model Evaluation Cycle

Conduct Evaluations

Perform state-of-the-art model evaluations.

Design Evaluations

Tailor methods to model and risk.

Open-Ended Testing

Explore unexpected behaviors.

Gather Information

Collect model-independent data.

Inform Design

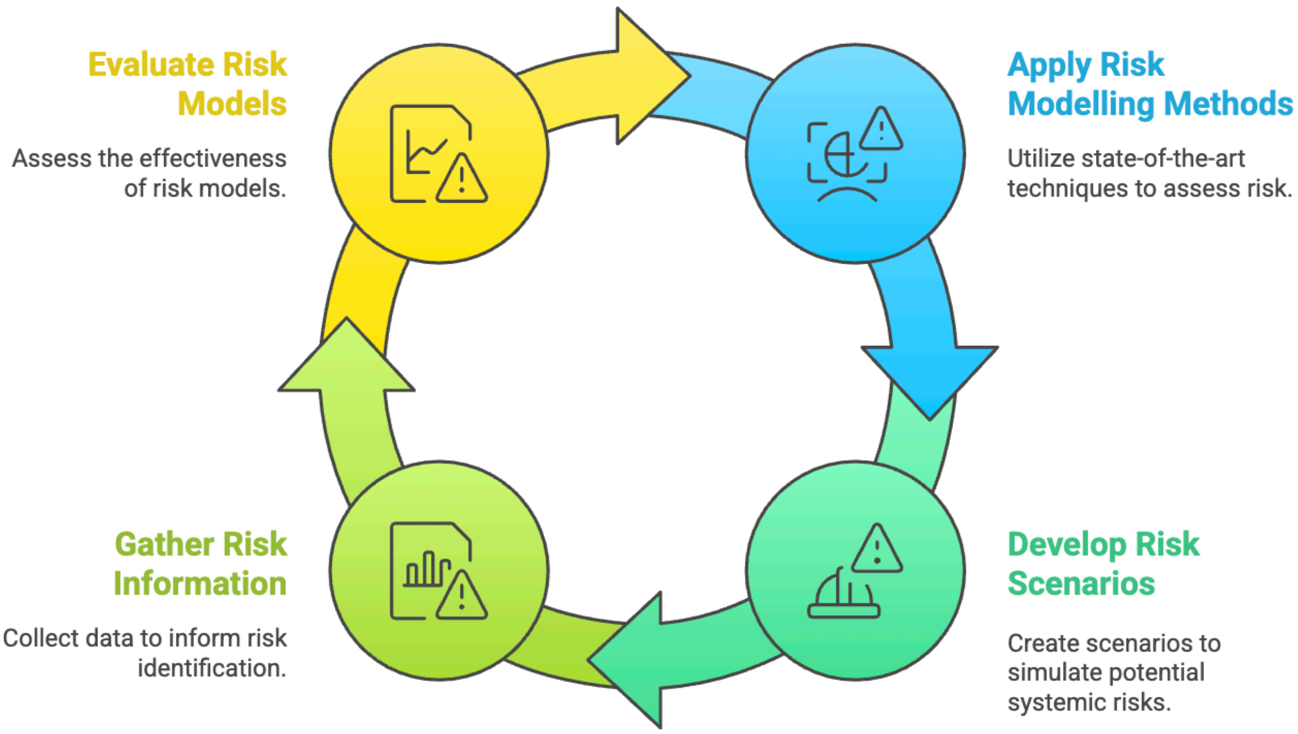
Use data to refine evaluations.

Measure 3.3 Systemic risk modelling

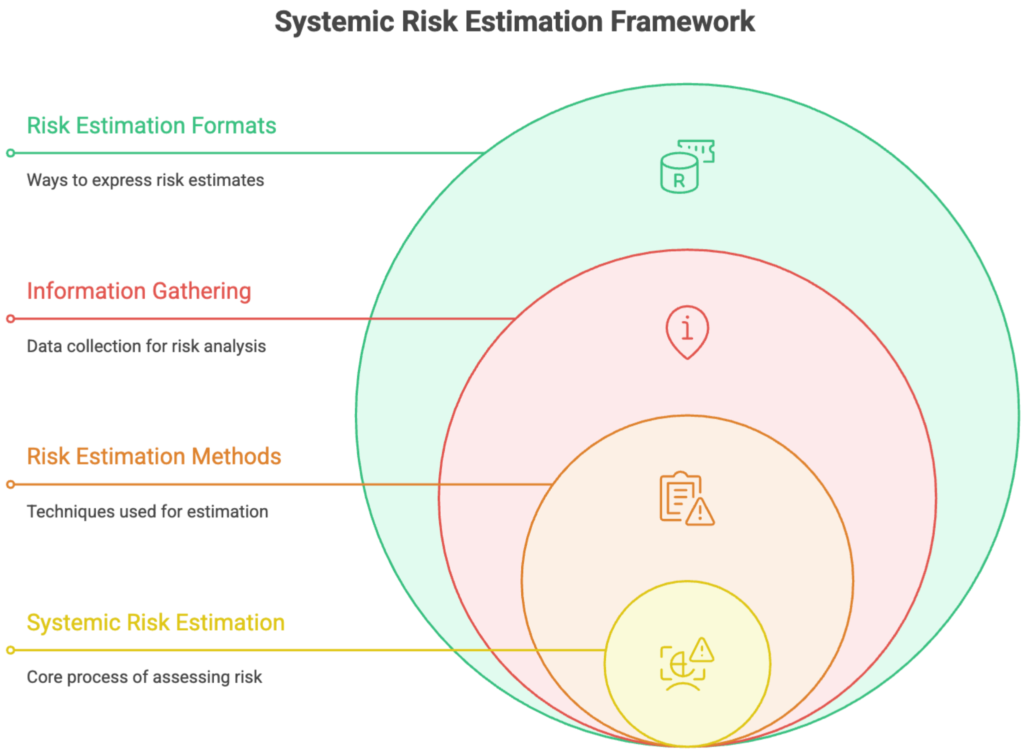
Conduct systemic risk modelling is necessary to carry out and identify the systemic risk using :

(1) at least state-of-the-art risk modelling methods;
(2) the systemic risk scenarios developed; and
(3) the systemic risk identification information gathered
(4) model evaluations of systemic risk

Systemic Risk Modelling Cycle

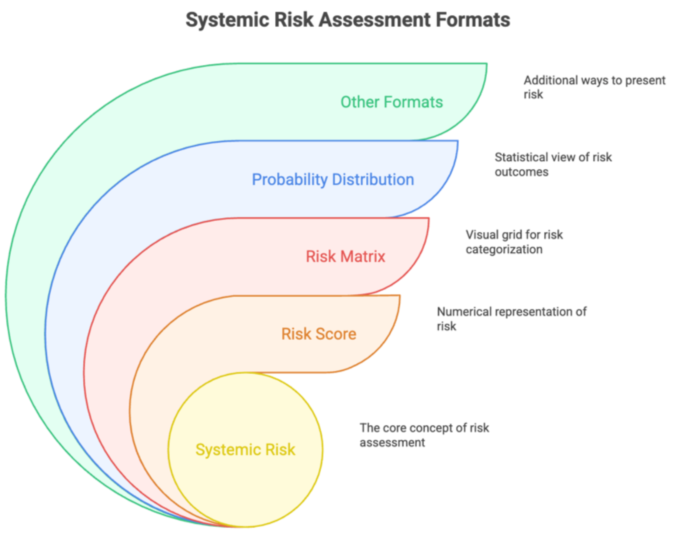

Evaluate Risk Models
Assess the effectiveness of risk models.
Apply Risk Modelling Methods
Utilize state-of-the-art techniques to assess risk.
Gather Risk Information
Collect data to inform risk identification.
Develop Risk Scenarios
Create scenarios to simulate potential systemic risks.

Measure 3.4 Systemic risk estimation


Systemic Risk Estimation Framework
Risk Estimation Formats
Ways to express risk estimates
Information Gathering
Data collection for risk analysis
Risk Estimation Methods
Techniques used for estimation
Systemic Risk Estimation
Core process of assessing risk

It is necessary to estimate the probability and severity of harm for the systemic risk, using at least state-of-the-art risk estimation methods and take into account at least the information gathered in relation to Systemic Risk Identification, Systemic Risk Analysis and any Serious Incident Reporting.

Estimates of systemic risk will be expressed as a risk score, risk matrix, probability distribution, or in other adequate formats, standard to industry and may be quantitative, semi-quantitative, and/or qualitative.


Systemic Risk Assessment Formats
Other Formats
Additional ways to present risk
Probability Distribution
Statistical view of risk outcomes
Risk Matrix
Visual grid for risk categorization
Risk Score
Numerical representation of risk
Systemic Risk
The core concept of risk assessment

Examples of such estimates of systemic risks are:
(1) a qualitative systemic risk score (e.g. “moderate” or “critical”);
(2) a qualitative systemic risk matrix (e.g. “probability: unlikely” x “impact: high”); and/or
(3) a quantitative systemic risk matrix (e.g. “X-Y%” x “X-Y EUR damage”).

Measure 3.5 Post-market monitoring

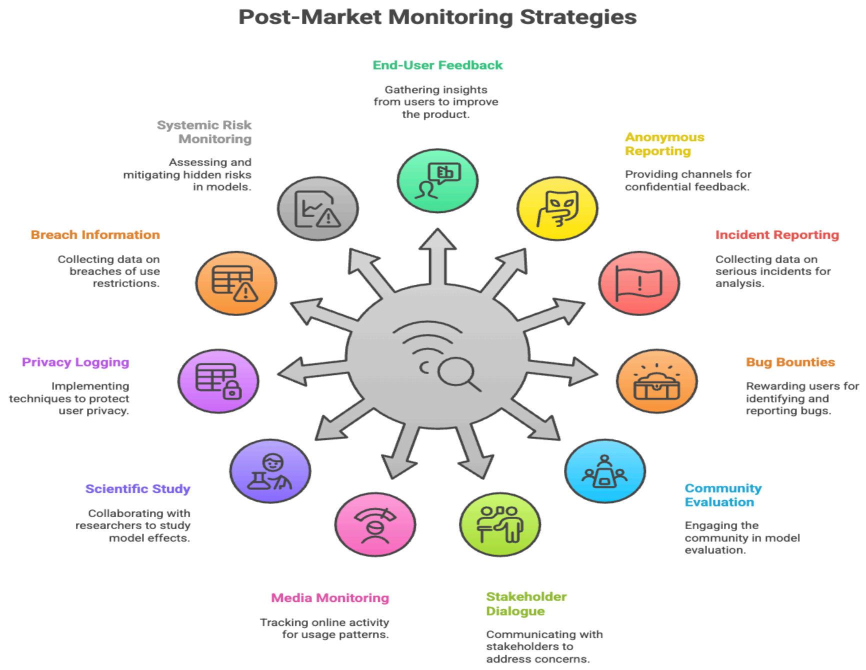
Appropriate post-market monitoring must be conducted to gather information relevant to assessing whether the systemic risk could be determined to not be acceptable and to inform whether a Model Report update is necessary and it will be necessary to use best efforts to conduct post-market monitoring to gather information relevant to producing estimates of timelines. To these ends, post-market monitoring will:

(1) gather information about the model’s capabilities, propensities, affordances, and/or effects;

(2) take into account the exemplary methods listed below; and
(3) if own system models are provided and/or deployed this would include monitoring of internal the models.

The following are examples of post-market monitoring methods:

(1) collecting end-user feedback;
(2) providing (anonymous) reporting channels;
(3) providing (serious) incident reporting forms;
(4) providing bug bounties;
(5) establishing community-driven model evaluations and public leaderboards;
(6) conducting frequent dialogues with affected stakeholders;
(7) monitoring software repositories, malware alerts, public forums, &/or social media usage patterns;
(8) supporting the scientific study of the model's capabilities, propensities, affordances, and/or effects in collaboration with academia, civil society, regulators, and/or independent researchers;
(9) implementing privacy-preserving logging & metadata analysis techniques of model inputs & outputs using systems such as watermarks, metadata, and/or other state-of-the-art provenance techniques;
(10) collecting relevant information about breaches of the model's use restrictions and subsequent incidents arising from such breaches; and/or
(11) monitoring aspects of models that are relevant to assess & mitigate systemic risk & are not transparent to third parties, such as hidden chains-of-thought for models for non-public parameters.


Post-Market Monitoring Strategies
End-User Feedback
Gathering insights from users to improve the product.
Anonymous Reporting
Providing channels for confidential feedback.
Incident Reporting
Collecting data on serious incidents for analysis.
Bug Bounties
Rewarding users for identifying and reporting bugs.
Community Evaluation
Engaging the community in model evaluation.
Stakeholder Dialogue
Communicating with stakeholders to address concerns.
Media Monitoring
Tracking online activity for usage patterns.
Scientific Study
Collaborating with researchers to study model effects.
Privacy Logging
Implementing techniques to protect user privacy.
Breach Information
Collecting data on breaches of use restrictions.
Systemic Risk Monitoring
Assessing and mitigating hidden risks in models.

Unless the model is considered a similarly safe or safer model with regard to the same systemic risk, the post-market monitoring process under the Code envisages that there will be an adequate number of independent external evaluators able to assess the most capable model version(s) for:

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

It is envisaged that
(i) the number of such evaluators, the selection criteria, and security measures will differ for points (1), (2), and (3) above;

(ii) Providers will publish suitable criteria for assessing applications;
(ii) Providers will allow access to a model through an API, on- premise access, access via provider-provided hardware, or by making the model parameters publicly available for download, as appropriate;

Providers must only use the evaluation results from independent external evaluators to assess and mitigate systemic risk from the model and the Code suggests that providers should refrain from training their models on the inputs and/or outputs from such test runs without express permission from the evaluators. This is a rather strange provision and there is no explanation for this in the Code documentation.
The Code originally provided for GPAISRs to be monitored after the relevant AI was officially retired although it is good to see that this provision has now been removed.

Additionally, Signatories will not take any legal or technical retaliation against the independent external evaluators as a consequence of their testing and/or publication of findings as long as the evaluators:

(1) do not intentionally disrupt model availability through the testing, unless expressly permitted;
(2) do not intentionally access, modify, and/or use sensitive or confidential user data in violation of Union law, and if evaluators do access such data, collect only what is necessary, refrain from disseminating it, and delete it as soon as legally feasible;
(3) do not intentionally use their access for activities that pose a significant risk to public safety and security;
(4) do not use findings to threaten Signatories, users, or other actors in the value chain, provided that disclosure under pre-agreed policies and timelines will not be counted as such coercion; and
(5) adhere to the Signatory's publicly available procedure for responsible vulnerability disclosure, which will specify at least that the Signatory cannot delay or block publication for more than 30 business days from the date that the Signatory is made aware of the findings, unless a longer timeline is exceptionally necessary such as if disclosure of the findings would materially increase the systemic risk.

Signatories that are SMEs or SMCs may contact the AI Office, which may provide support or resources to facilitate adherence to this Measure.

COMMITMENT 4
SYSTEMIC RISK ACCEPTANCE DETERMINATION

This is a commitment to specifying systemic risk acceptance criteria and determining whether the systemic risks stemming from the model are acceptable and a commitment to decide whether or not to proceed with the development, the making available on the market, and/or the use of the model based on the systemic risk acceptance determination.

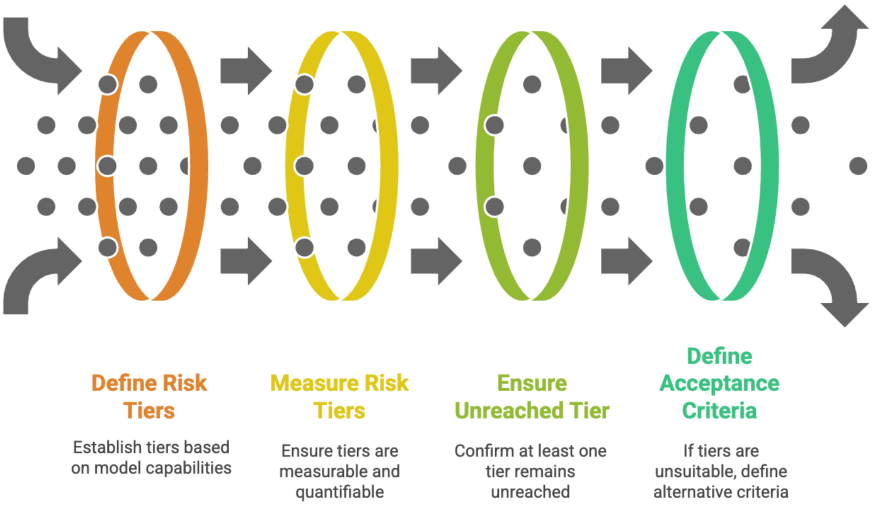
Measure 4.1 Systemic Risk Acceptance Criteria and Acceptance Determination

There will have been described and justified in the Framework how the assessments will determine whether the systemic risks stemming from the model are acceptable.

For each identified systemic risk, it will be necessary to, at least:

(a) define appropriate systemic risk tiers that:
  (i) are defined in terms of model capabilities, and may additionally incorporate model propensities, risk estimates, and/or other suitable metrics;
  (ii) are measurable; and
  (iii) comprise at least one systemic risk tier that has not been reached by the model; or
(b) define other appropriate systemic risk acceptance criteria, if systemic risk tiers are not suitable for the systemic risk and the systemic risk is not a specified systemic risk[65].

Systemic Risk Management Process


Define Risk Tiers
Establish tiers based on model capabilities
Measure Risk Tiers
Ensure tiers are measurable and quantifiable
Ensure Unreached Tier
Confirm at least one tier remains unreached
Define Acceptance Criteria
If tiers are unsuitable, define alternative criteria

It will also be necessary to describe how each of these tiers are used and/or other criteria to determine whether each identified systemic risk and the overall systemic risk are acceptable as well as to justify how the use of these tiers and/or other criteria ensures that each identified systemic risk and the overall systemic risk are acceptable.

The systemic risk acceptance criteria for each identified systemic risk must then be applied and must incorporate a safety margin to determine whether (i) each identified systemic risk and (ii) the overall systemic risk are acceptable having regard to the information gathered via systemic risk identification and analysis and any other relevant factors.

[65] see Appendix 1.4

The safety margins must:

(1) be appropriate for the systemic risk; and

(2) take into account potential limitations, changes, and uncertainties of:

(a) systemic risk sources (e.g. capability improvements after the time of assessment);

(b) systemic risk assessments (e.g. under-elicitation of model evaluations or historical accuracy of similar assessments); and

(c) the effectiveness of safety and security mitigations (e.g. mitigations being circumvented, deactivated, or subverted).

Safety Margin Considerations

Safety Margins
Core safety measures to mitigate risk
Systemic Risk
Overall risk level requiring safety measures
Risk Sources
Factors contributing to systemic risk
Risk Assessments
Evaluations of potential risks
Mitigation Effectiveness
Reliability of safety measures

Measure 4.2
Proceeding or Not Proceeding based on Systemic Risk Acceptance Determination

Only if the systemic risks stemming from the model are determined to be acceptable according to the above criteria is it permitted to proceed with the development, the making available on the market, and/or the use of the model.

If the systemic risks stemming from the model are not determined to be acceptable or will be unacceptable within a reasonably foreseeable period, then it will be necessary to take appropriate measures to ensure the systemic risks stemming from the model are and will remain acceptable prior to proceeding and that the unacceptable risk identified is mitigated so that it does not occur. In particular, unless this can be achieved so that the risk remains acceptable:

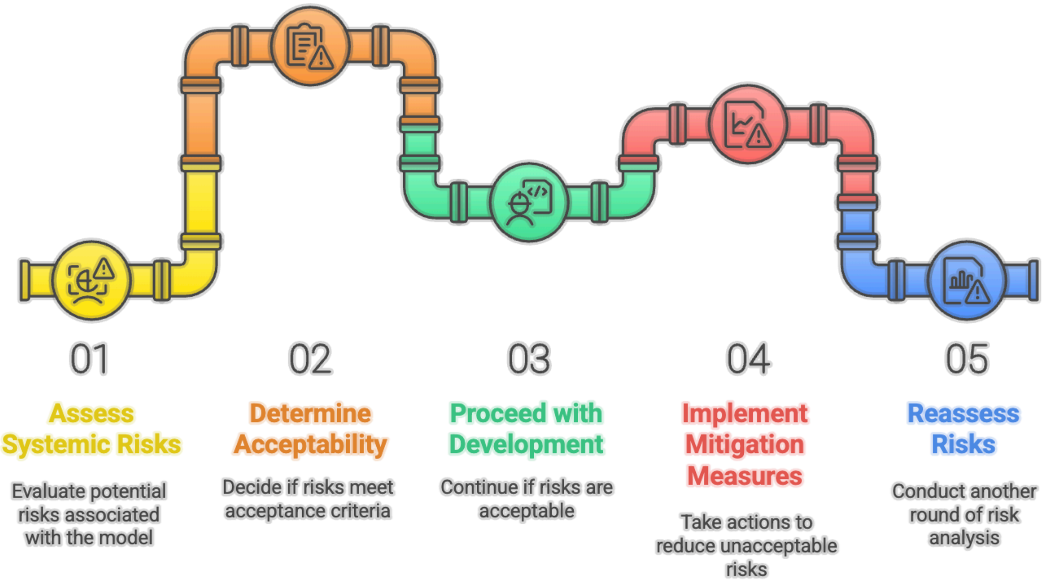
(1) The model must not be made available on the market, or there must be restriction on the making available on the market of the model by suitably adjusted license conditions or usage restrictions, or the model must be withdraw or recalled as necessary;

(2) There must be implemented immediate safety and/or security mitigations[66]; and

(3) There must be conducted another round of systemic risk identification, systemic risk analysis and systemic risk acceptance determinations.

[66] See Commitments 5 and 6

Systemic Risk Management Process



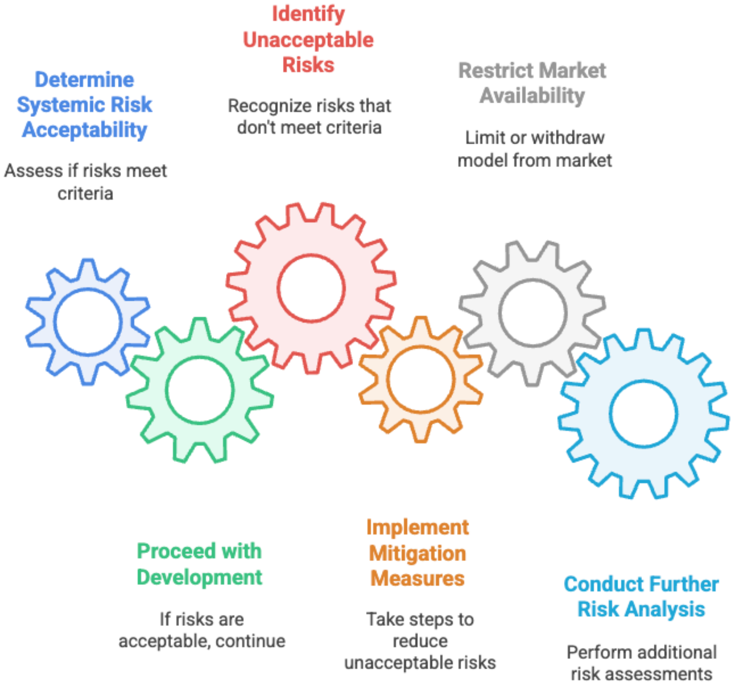
01 **Assess Systemic Risks**
Evaluate potential risks associated with the model

02 **Determine Acceptability**
Decide if risks meet acceptance criteria

03 **Proceed with Development**
Continue if risks are acceptable

04 **Implement Mitigation Measures**
Take actions to reduce unacceptable risks

05 **Reassess Risks**
Conduct another round of risk analysis



Identify Unacceptable Risks
Recognize risks that don't meet criteria

Determine Systemic Risk Acceptability
Assess if risks meet criteria

Restrict Market Availability
Limit or withdraw model from market

Proceed with Development
If risks are acceptable, continue

Implement Mitigation Measures
Take steps to reduce unacceptable risks

Conduct Further Risk Analysis
Perform additional risk assessments

COMMITMENT 5
SAFETY MITIGATIONS

A commitment to implement appropriate safety mitigations along the entire model lifecycle to ensure the systemic risks stemming from the model are acceptable.

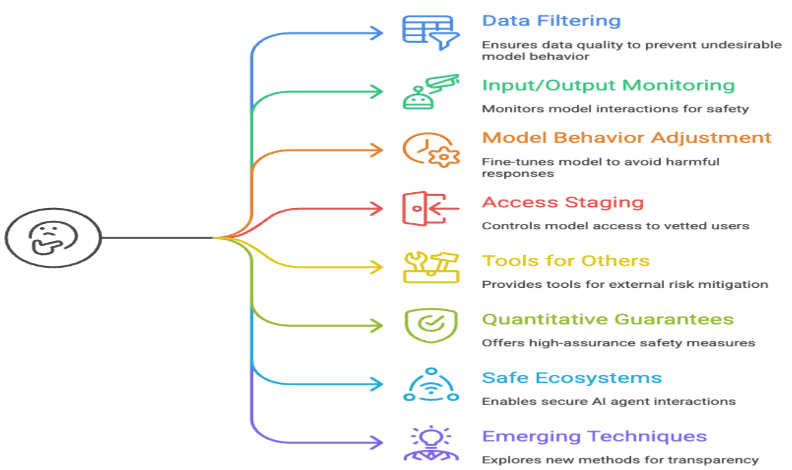
Measure 5.1 Appropriate Safety Mitigations

It is necessary to ensure that safety mitigations are appropriate, including being sufficiently robust under adversarial pressure (e.g. fine-tuning attacks or jailbreaking), taking into account the model's release and distribution strategy.

Safety mitigations examples include (but not exclusively):

(1) filtering and cleaning training data[67];
(2) monitoring and filtering the model's inputs and/or outputs;
(3) changing the model behaviour in the interests of safety, such as fine-tuning the model to refuse certain requests or provide unhelpful responses;
(4) staging the access to the model[68], and/or not making the model parameters publicly available for download initially;
(5) offering tools for other actors to use to mitigate the systemic risks;
(6) techniques that provide high-assurance quantitative safety guarantees concerning the model's behaviour;
(7) techniques to enable safe ecosystems of AI agents[69]; and/or
(8) other emerging safety mitigations, such as for achieving transparency into chain-of-thought reasoning or defending against a model's ability to subvert its other safety mitigations.

Which safety mitigation strategy should be implemented?


Data Filtering
Ensures data quality to prevent undesirable model behavior
Input/Output Monitoring
Monitors model interactions for safety
Model Behavior Adjustment
Fine-tunes model to avoid harmful responses
Access Staging
Controls model access to vetted users
Tools for Others
Provides tools for external risk mitigation
Quantitative Guarantees
Offers high-assurance safety measures
Safe Ecosystems
Enables secure AI agent interactions
Emerging Techniques
Explores new methods for transparency

[67] e.g. data that might result in undesirable model propensities such as unfaithful chain-of-thought traces
[68] e.g. by limiting API access to vetted users, gradually expanding access based on post-market monitoring
[69] such as model identifications, specialised communication protocols, or incident monitoring tools

COMMITMENT 6
SECURITY MITIGATIONS[70]

1. **A commitment to implementing an adequate level of cybersecurity protection for all models and their physical infrastructure along the entire model lifecycle.**
2. **A commitment to ensure the systemic risks arising from unauthorised releases, unauthorised access, and/or model theft are acceptable.**

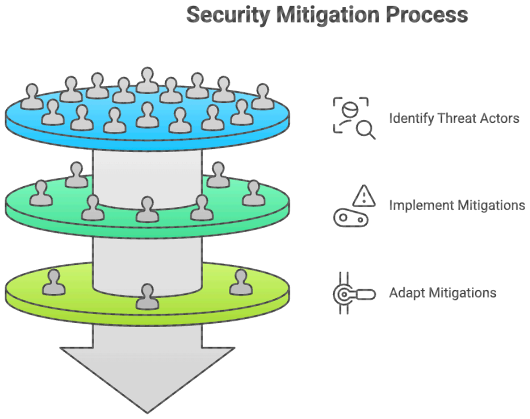
NB: A model is exempt from this Commitment if the model's capabilities are inferior to the capabilities of at least one model for which the parameters are publicly available for download.
It is necessary to implement these security mitigations for a model until its parameters are made publicly available for download or securely deleted.

Measure 6.1 Security Goal

It is necessary to create and define a goal that specifies the threat actors that their security mitigations[71] are intended to protect against ("Security Goal"), including non-state external threats, insider threats, and other expected threat actors, taking into account at least the current and expected capabilities of their models.

Measure 6.2 Implement appropriate security mitigations

Following the creation of the Security Goal, it is necessary to implement appropriate security mitigations to meet the Security Goal, including the Appendix 4 security mitigations. Where there is deviation[72] from any of those security mitigations, then it is necessary that implementation of alternative security mitigations that achieve the respective mitigation objectives occurs. The implementation of the required security mitigations may be staged appropriately in line with the increase in model capabilities along the entire model lifecycle.



[70] Article 55(1), and recitals 114 and 115 AI Act
[71] Including measures such as defences against adversarial fine-tuning, adversarial jailbreaking, data filtering and poisoning, and will include transparency tools, downstream risk controls, refusal fine-tuning and guardrails etc. It will also include latest state of art cybersecurity measures to combat unauthorised releases, access, changes and theft and should address all known threat actors as well as foreseeable ones.
[72] In some cases, for example, the organisational context and digital infrastructure will require deviation and this provisions ensures the achievement of the respective mitigation objectives.

COMMITMENT 7
SAFETY AND SECURITY MODEL REPORTS[73]

1. **A commitment to reporting to the AI Office information about their model and their systemic risk assessment and mitigation processes and measures by creating a Safety and Security Model Report ("Model Report") before placing a model on the market.**
2. **A commitment to keeping the Model Report up-to-date and notifying the AI Office of their Model Report as updated.**

NB: The original test, laid down in draft version required document retention for 12 months following retirement. In this final version the technical documentation must be retained for 10 years from the time after the model is placed on the market. For serious incidents this period is extended for 5 years to be measured from the later of the date of the documentation or the date of the serious incident.

If there has already been provided relevant information to the AI Office in other reports and/or notifications, reference those reports and/or notifications may be made in the Model Report, and a single Model Report may be created for several models if the systemic risk assessment and mitigation processes and measures for one model cannot be understood without reference to the other model(s).

Note: SMEs or SMCs may reduce the level of detail in their Model Report to the extent necessary to reflect size and capacity constraints.

Measure 7.1 Model description and behaviour

The Model Report must provide:

(1) a high-level description of the model's architecture, capabilities, propensities, and affordances, and how the model has been developed, including its training method and data, as well as how these differ from other models that the reporting party has made available on the market;

(2) a description of how the model has been used and is expected to be used, including its use in the development, oversight, and/or evaluation of models;

(3) a description of the model versions that are going to be made or are currently made available on the market and/or used, including differences in systemic risk mitigations and systemic risks; and

(4) a specification (e.g. via valid hyperlinks) of how Signatories intend the model to operate (often known as a "model specification"), including by:

   (a) specifying the principles that the model is intended to follow;

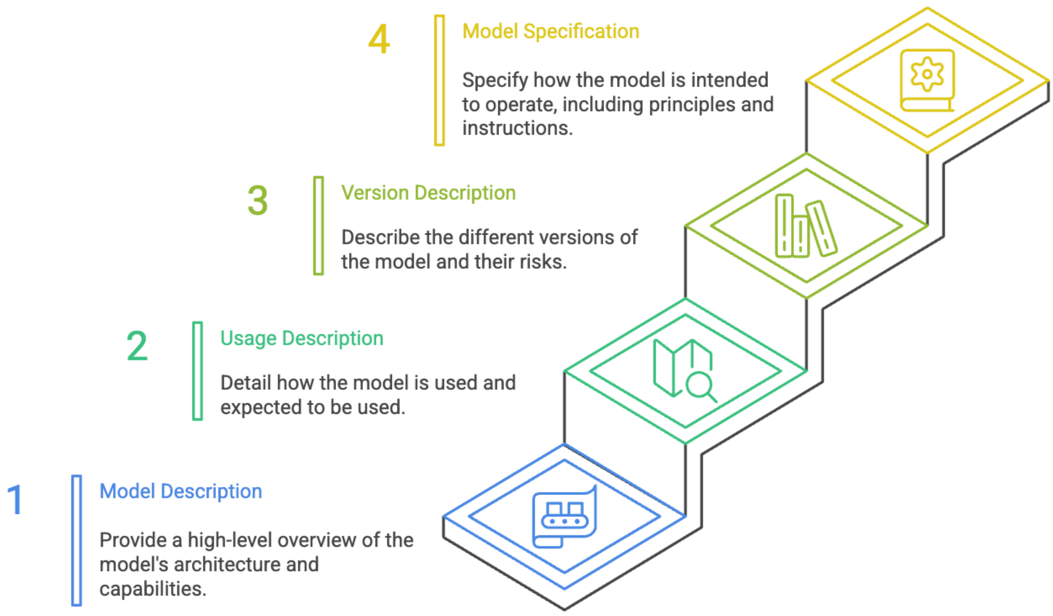
   (b) stating how the model is intended to prioritise different kinds of principles and instructions;

   (c) listing topics on which the model is intended to refuse instructions; and

   (d) providing the system prompt.

[73] Articles 55(1) and 56(5) AI Act

Model Report Development



4 Model Specification

Specify how the model is intended to operate, including principles and instructions.

3 Version Description

Describe the different versions of the model and their risks.

2 Usage Description

Detail how the model is used and expected to be used.

1 Model Description

Provide a high-level overview of the model's architecture and capabilities.

Measure 7.2 Reasons for proceeding

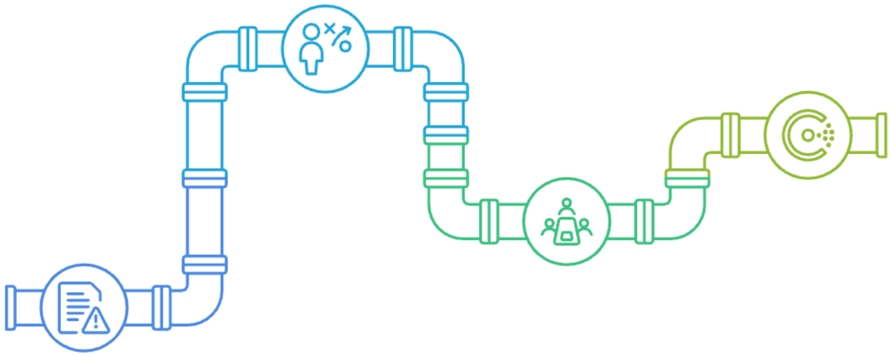
The Model Report must provide:

(1) a detailed justification for why the systemic risks stemming from the model are acceptable, including details of the safety margins incorporated (pursuant to Measure 4.1);

(2) the reasonably foreseeable conditions under which the justification in point (1) would no longer hold; and

(3) a description of how the decision to proceed with the development, making available on the market, and/or use was made, including whether input from external actors informed such a decision, and whether and how input from independent external evaluators pursuant to Appendix 3.5 informed such a decision.

Model Report Signatory Responsibilities



01

Provide Justification for Systemic Risks

Signatories must justify why the model's systemic risks are acceptable, including safety margins.

02

Identify Conditions for Justification Failure

Signatories must identify conditions under which the risk justification would no longer hold.

03

Describe Decision-Making Process

Signatories must describe how the decision to proceed with the model was made.

04

Include External Input

Signatories must include input from external actors and evaluators in the decision-making process.

Measure 7.3 Documentation of systemic risk identification, analysis, and mitigation

The Model Report should provide:

(1) a description of the results of their systemic risk identification and analysis and any information relevant to understanding them including:

(a) a description of their systemic risk identification process for risks belonging to the types of risks in Appendix 1.1;

(b) explanations of uncertainties and assumptions about how the model would be used and integrated into AI systems;

(c) a description of the results of their systemic risk modelling for the systemic risks;

(d) a description of the systemic risks stemming from the model and a justification therefor, including:

(i) the systemic risk estimates; and

(ii) a comparison between systemic risks with safety and security mitigations implemented and with the model fully elicited (pursuant to Appendix 3.2);

(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. It is also necessary to provide a sufficiently large number of random samples of inputs and outputs from a relevant model evaluation *if subsequently* requested by the AI Office;

(g) a description of the access and other resources provided to:
(i) internal model evaluation teams (pursuant to Appendix 3.4); and
(ii) independent external evaluators (pursuant to Appendix 3.5)[74]; and

(h) if they make use of the “similarly safe or safer model” concept pursuant to Appendix 2, provide a justification of how the criteria for “safe reference model” (pursuant to Appendix 2.1) and the criteria for “similarly safe or safer model” (pursuant to Appendix 2.2) are fulfilled.

(2) a description of:
(a) all safety mitigations implemented (pursuant to Commitment 5);
(b) how they fulfil the requirements of Measure 5.1; and
(c) their limitations (e.g. if training on examples of undesirable model behaviour makes identifying future instances of such behaviour more difficult).

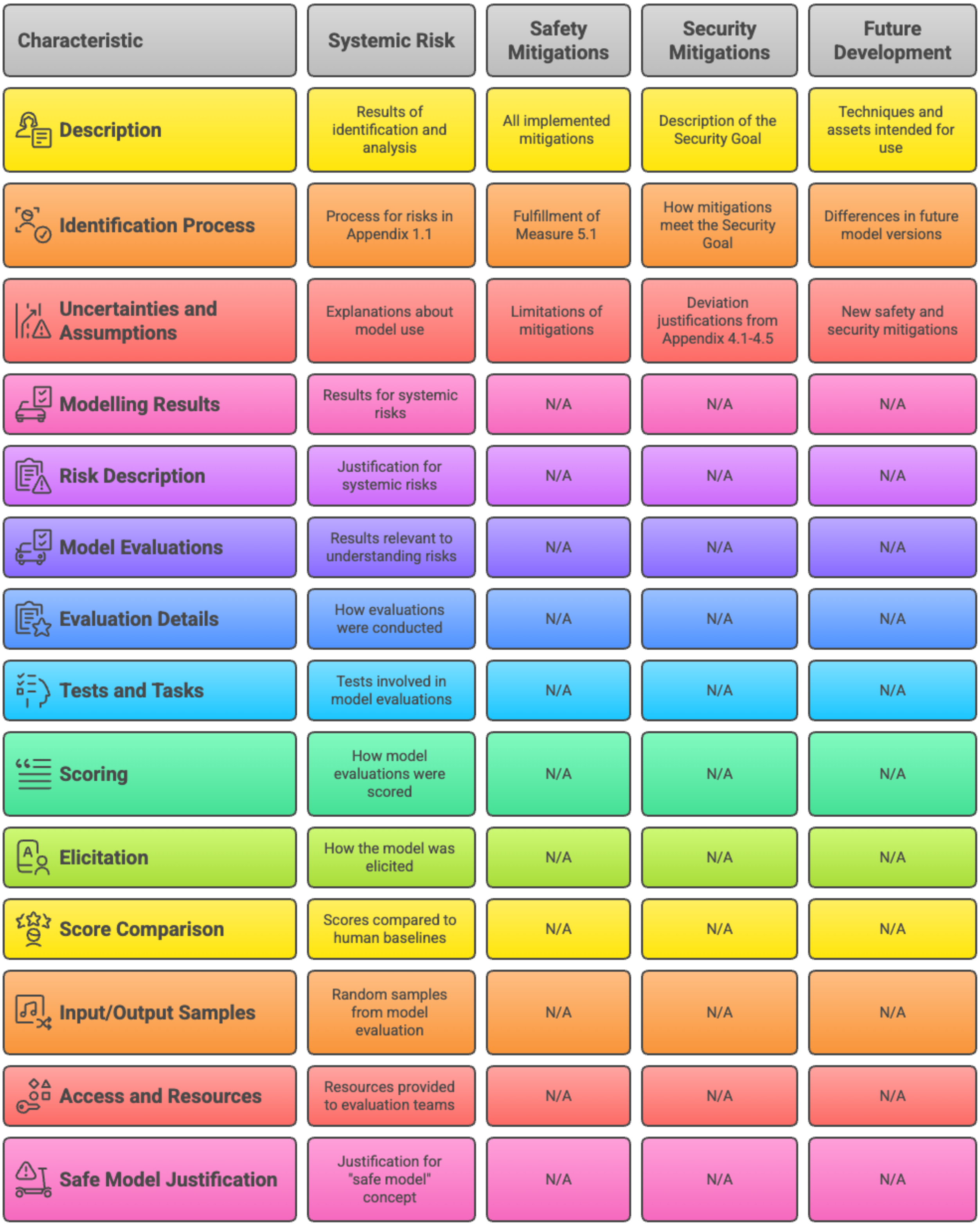
(3) a description of:
(a) the Security Goal;
(b) all security mitigations implemented;
(c) how the mitigations meet the Security Goal, including the extent to which they align with relevant international standards or other relevant guidance[75]; and
(d) if there has been a deviation from a listed security mitigation in one (or more) of Appendices 4.1 to 4.5, points (a), then it is necessary to provide a justification for how the alternative security mitigations they have implemented achieve the respective mitigation objectives; and

(4) a high-level description of:
(a) the techniques and assets they intend to use to further develop the model over the next six months, including through the use of other AI models and/or AI systems;
(b) how such future versions and more advanced models may differ from the Signatory’s current ones, in terms of capabilities and propensities; and
(c) any new or materially updated safety and security mitigations that they intend to implement for such models.

[74] As an alternative to the point (ii), it is possible to procure independent external evaluators to provide the requisite information directly to the AI Office at the same time that the Signatory supplies its Model Report to the AI Office

[75] (such as the RAND Securing AI Model Weights report. It should be noted that the original test, laid down in draft version 3 specified the RAND SL3 test developed by the RAND Corporation in a 2024 research report and set out the level of security needed to protect model weights against well-resourced non-state actors. The final version no longer applies this baseline and requires *latest state of art* defences.

Model Report Requirements



Characteristic	Systemic Risk	Safety Mitigations	Security Mitigations	Future Development
Description	Results of identification and analysis	All implemented mitigations	Description of the Security Goal	Techniques and assets intended for use
Identification Process	Process for risks in Appendix 1.1	Fulfillment of Measure 5.1	How mitigations meet the Security Goal	Differences in future model versions
Uncertainties and Assumptions	Explanations about model use	Limitations of mitigations	Deviation justifications from Appendix 4.1-4.5	New safety and security mitigations
Modelling Results	Results for systemic risks	N/A	N/A	N/A
Risk Description	Justification for systemic risks	N/A	N/A	N/A
Model Evaluations	Results relevant to understanding risks	N/A	N/A	N/A
Evaluation Details	How evaluations were conducted	N/A	N/A	N/A
Tests and Tasks	Tests involved in model evaluations	N/A	N/A	N/A
Scoring	How model evaluations were scored	N/A	N/A	N/A
Elicitation	How the model was elicited	N/A	N/A	N/A
Score Comparison	Scores compared to human baselines	N/A	N/A	N/A
Input/Output Samples	Random samples from model evaluation	N/A	N/A	N/A
Access and Resources	Resources provided to evaluation teams	N/A	N/A	N/A
Safe Model Justification	Justification for "safe model" concept	N/A	N/A	N/A

Measure 7.4 External reports

The Model Report should include any available reports[76] from:
a) independent external evaluators involved in model evaluations pursuant to Appendix 3.5; an
b) security reviews undertaken by an independent external party pursuant to Appendix 4.5.

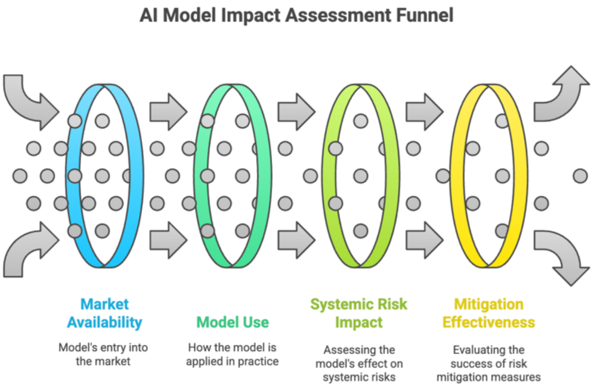
They are only required to the extent that they do not breach existing confidentiality[77] obligations. In addition, such reports are only required where external evaluators or parties to maintain control over the reports and can publish them without implicit endorsement of the AI providers[78].

If no independent external evaluator was involved in model evaluations pursuant to Appendix 3.5, then it is necessary to provide a justification of how the conditions in Appendix 3.5, first paragraph, points (1) or (2) were met.

If at least one independent external evaluator was involved in model evaluations pursuant to Appendix 3.5, then the Report must provide an explanation of the choice of evaluator based on the qualification criteria.

Measure 7.5 Identifying Material changes to the systemic risk landscape

The Model Report must also contain information that enables the AI Office to understand how the development of the model, its availability on the market and its use will impact on the systemic risk landscape and whether those impacts are material and also to determine whether the implementation of systemic risk assessment and mitigation measures and processes has been effective in controlling systemic risk and mitigating any identified risks.This might include such factors as:
(i) providing a description of scaling laws that suggest novel ways of improving model capabilities;
(ii) a summary of the characteristics of novel architectures that materially improve the state of the art in computational efficiency or model capabilities;
(iii) descriptions of information relevant to assessing the effectiveness of mitigations, for example if the model's chain-of-thought is less intelligible to humans;
(iv) description of training techniques that materially improve the efficiency or feasibility of distributed training.


AI Model Impact Assessment Funnel
Market Availability
Model's entry into the market
Model Use
How the model is applied in practice
Systemic Risk Impact
Assessing the model's effect on systemic risks
Mitigation Effectiveness
Evaluating the success of risk mitigation measures

[76] These additional reports can be provided by hyperlinks in the Model Report.
[77] including commercial confidentiality
[78] For example, if the AI provider retains intellectual property rights in the report then the external evaluators do not maintain control over the publication of their findings and the report does not have to be included.

Measure 7.6 Model Report updates

Model Report updates are required if there are reasonable grounds to believe that the justification from the model systemic risk assessment that systemic risk is acceptable has been materially undermined and these must contain changelogs.

This might be that one of the reasonably foreseeable conditions which were listed as potentially causing concerns over adequacy of safety margins has materialized. Alternatively, it may be the case that the model's capabilities, propensities, and/or affordances have changed or are expected to change materially in the very near future, whether through further post-training, access to additional tools, or increase in inference compute. Alternatively, the model's use and/or integrations into AI systems could have changed or might be expected to change materially or there could have been serious incidents and/or near misses involving the model or a similar model or developments might have occurred that materially undermine the external validity of model safety evaluations conducted or other reasons might have arisen that suggest that the systemic risk safety assessment previously conducted is materially inaccurate.

Model Report updates should be completed within a reasonable amount of time after the Signatory becomes aware of the grounds that necessitate an update,

Note: If a Model Report update is triggered by a deliberate change to a model and that change is made available on the market, the Model Report update and the underlying full systemic risk assessment and mitigation process need to be completed before the change is made available on the market.

Where any model is amongst the most capable models available on the market, it is necessary to provide the AI Office with an updated Model Report at least every six months, unless:
(1) the model's capabilities, propensities, and/or affordances have not changed since they have last provided the AI Office with the Model Report, or update thereof; or
(2) a more capable model is being placed on the market in less than a month and reports are made in respect of this more capable models; or
(3) the model is considered similarly safe or safer for each identified systemic risk[79].

Measure 7.7 Model Report notifications

Signatories will provide the AI Office with access to the Model Report (without redactions, unless they are required by national security laws) by the time they place a model on the market, e.g. through a publicly accessible link or through a sufficiently secure channel specified by the AI Office.

If a Model Report is updated, it is necessary to provide the AI Office with access to the updated Model Report within five business days of a confirmed update.

To facilitate the placing on the market of a model, it is possible to delay providing the AI Office with a Model Report, or an update thereof, by up to 15 business days, but only where acting in good faith.

[79] (pursuant to Appendix 2.2)

COMMITMENT 8
SYSTEMIC RISK RESPONSIBILITY ALLOCATION[80]

1. **A commitment to defining clear responsibilities for managing the systemic risks stemming from their models across all levels of the organization.**
2. **A commitment to allocating appropriate resources to actors who have been assigned responsibilities for managing systemic risk**
3. **A commitment to: promoting a healthy risk culture.**

Measure 8.1 Definition of clear responsibilities

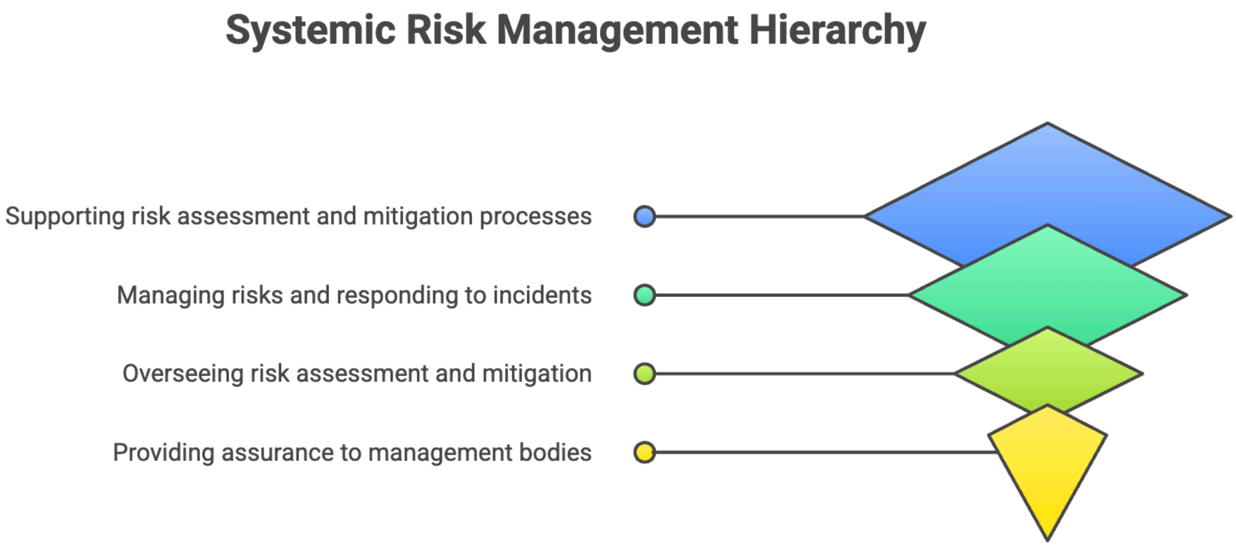
Under the Code it is now necessary to clearly define responsibilities for managing the systemic risks stemming from models across all levels of the organisation. This includes:

(1) **Systemic risk oversight:**
Overseeing the systemic risk assessment and mitigation processes and measures.

(2) **Systemic risk ownership:**
Managing systemic risks stemming from models, including the systemic risk assessment and mitigation processes and measures, and managing the response to serious incidents.

(3) **Systemic risk support and monitoring:**
Supporting and monitoring the systemic risk assessment and mitigation processes and measures.

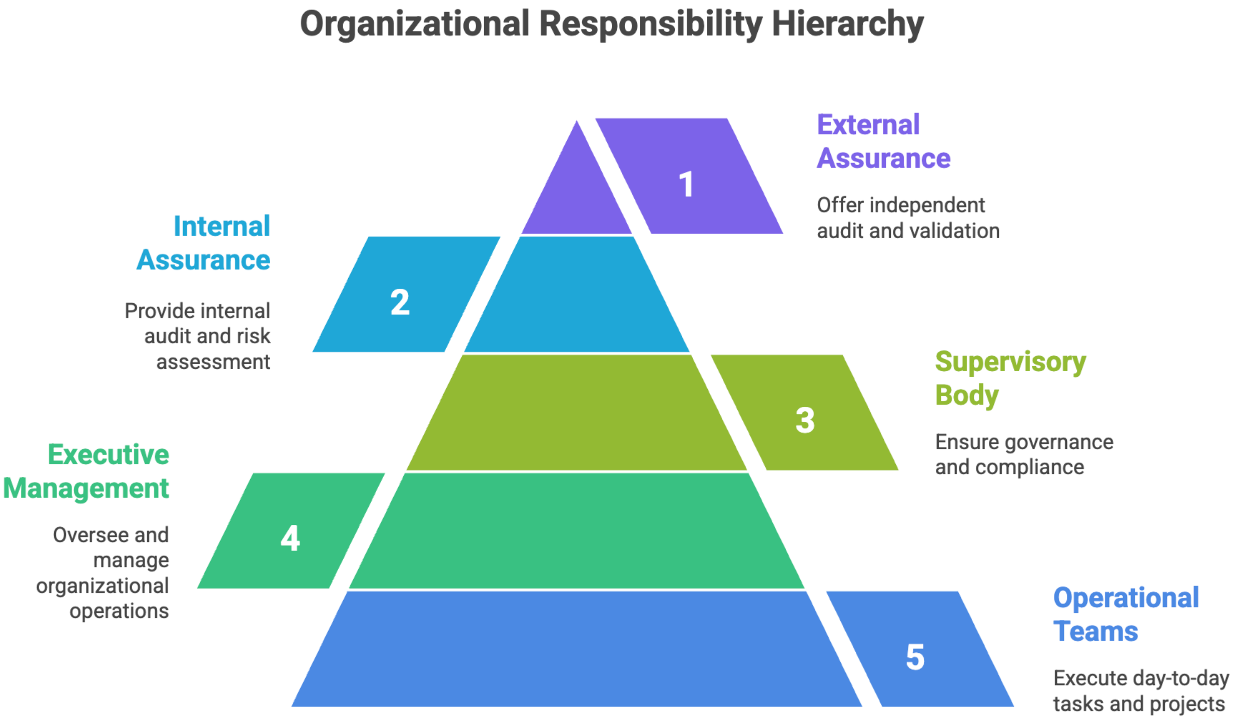
(4) **Systemic risk assurance:** Providing internal and, as appropriate, external assurance about the adequacy of the systemic risk assessment and mitigation processes and measures to the management body in its supervisory function or another suitable independent body (such as a council or board).


Systemic Risk Management Hierarchy
Supporting risk assessment and mitigation processes
Managing risks and responding to incidents
Overseeing risk assessment and mitigation
Providing assurance to management bodies

[80] Article 55(1) and recital 114 AI Act

These responsibilities must be allocated in a manner suitable for the governance structure and organisational complexity, across the following levels of an organisation:

(1) the management body in its supervisory function or another suitable independent body (such as a council or board);
(2) the management body in its executive function;
(3) relevant operational teams;
(4) if available, internal assurance providers (e.g. an internal audit function); and
(5) if available, external assurance providers (e.g. third-party auditors).


Organizational Responsibility Hierarchy
1
External Assurance
Offer independent audit and validation
2
Internal Assurance
Provide internal audit and risk assessment
3
Supervisory Body
Ensure governance and compliance
4
Executive Management
Oversee and manage organizational operations
5
Operational Teams
Execute day-to-day tasks and projects

Organisational responsibility obligations are deemed fulfilled if the following is adhered to (subject to this being appropriate for the systemic risks from the models:

(1) **Systemic risk oversight:**
The responsibility for overseeing the systemic risk management processes and measures has been assigned to:
(i) a specific committee of the management body in its supervisory function (e.g. a risk committee or audit committee)[81]; or
(ii) one or multiple suitable independent bodies (such as councils or boards).

(2) **Systemic risk ownership:**
The responsibility for managing systemic risks from models has been assigned to:
(i) suitable members of the management body in its executive function who are also responsible for relevant core business activities that may give rise to systemic risk[82]; or
(ii) Lower-level responsibilities have been assigned to operational managers[83] who oversee

[81] For SMEs or SMCs, this responsibility may be primarily assigned to an individual member of the management body in its supervisory function.
[82] such as research and product development (e.g. Head of Research or Head of Product)
[83] e.g. specific research domains or specific products

parts of the systemic-risk-producing business activities by members of the management body[84].

|

(3) **Systemic risk support and monitoring:**
The responsibility for supporting and monitoring the systemic risk management processes and measures, including conducting risk assessments, has been assigned to at least one member of the management body in its executive function[85] [86].
NB: This member(s) must not also be responsible for the Signatory's core business activities that may produce systemic risk (e.g. research and product development).

(4) **Systemic risk assurance:**
In most cases, the responsibility for providing assurance about the adequacy of the systemic risk assessment and mitigation processes and associated measures must be made to the management body in its supervisory function[87].
Where appropriate, this responsibility may be passed to another suitable independent body (such as a Security & Safety council or a Security & Safety board). This council or board may assign this responsibility to a relevant party[88] where appropriate provided that that party *is supported by an internal audit function, or equivalent, and external assurance as appropriate and subject to* internal assurance activities as are appropriate being included.
Note: SMEs or SMCs are permitted to assign the responsibility to the management body in its supervisory function providing that it periodically assesses the model's systemic risk assessment and mitigation processes and measures (for example by approving the internal Framework assessment).

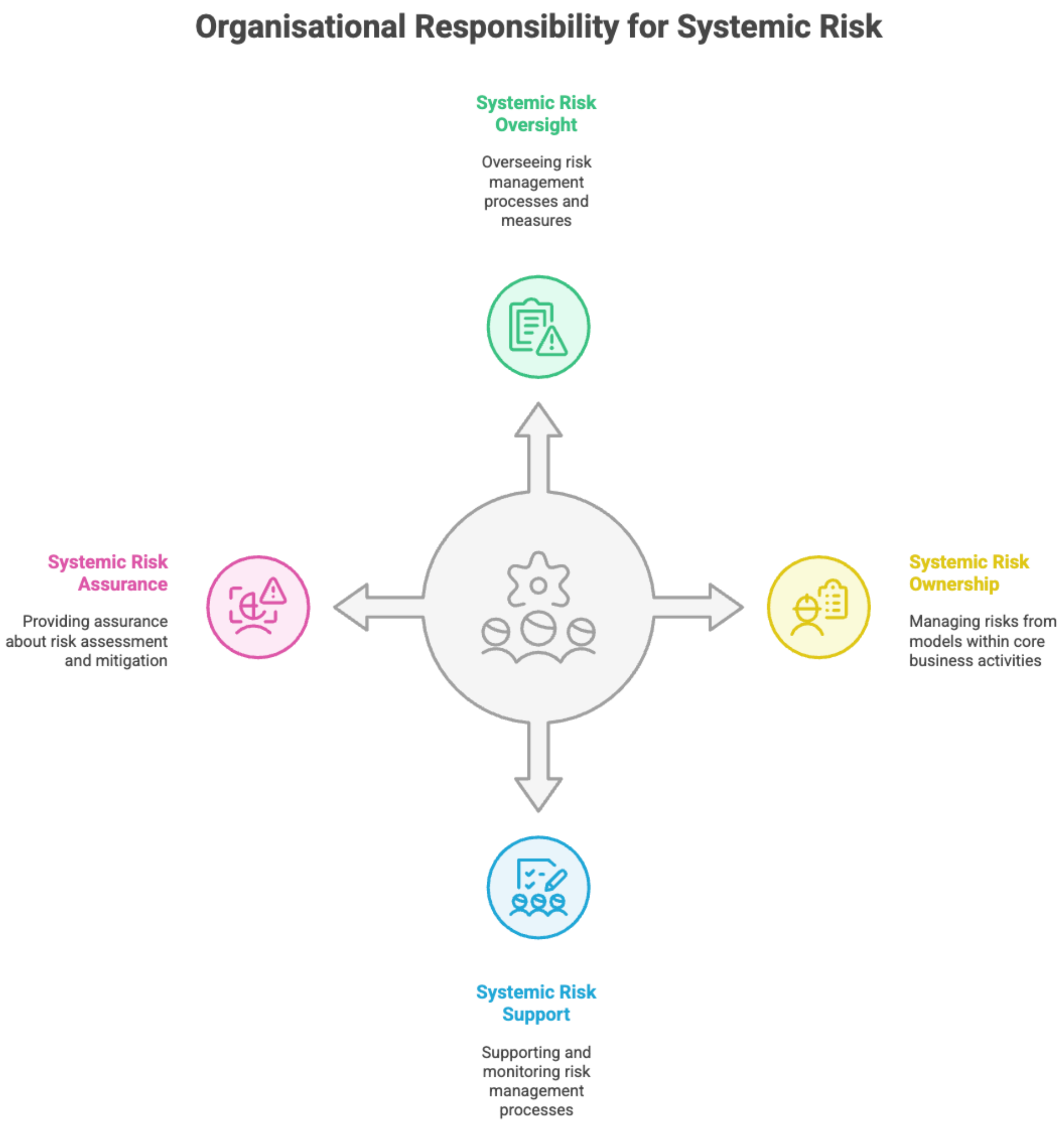
[84]

[85] (e.g. a Chief Risk Officer or a Vice President, Safety & Security Framework)

[86] For SMEs or SMCs, there is at least one individual in the management body in its executive function tasked with supporting and monitoring the systemic risk assessment and mitigation processes and measures.

[87] i.e. the Board

[88] for example, a Chief Audit Executive, a Head of Internal Audit, or a relevant sub-committee



Organisational Responsibility for Systemic Risk

Systemic Risk Oversight

Overseeing risk management processes and measures

Systemic Risk Assurance

Providing assurance about risk assessment and mitigation

Systemic Risk Ownership

Managing risks from models within core business activities

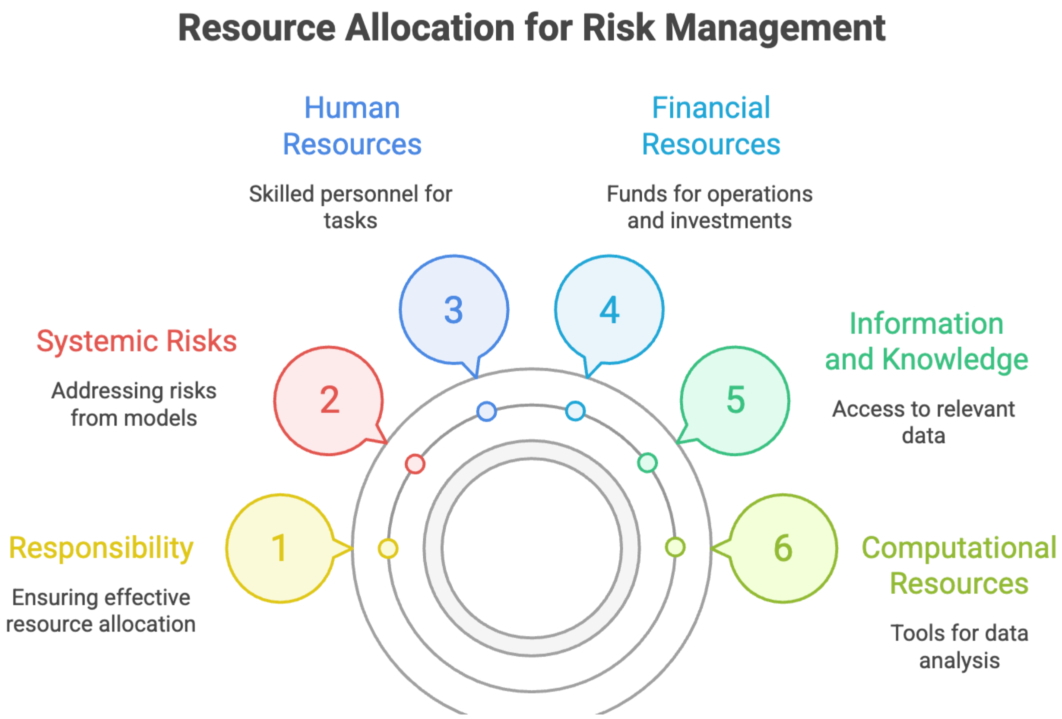
Systemic Risk Support

Supporting and monitoring risk management processes

Measure 8.2 Allocation of appropriate resources

It is necessary to ensure that management bodies oversee the allocation of resources to those who have been assigned responsibilities. These resources must be appropriate for the systemic risks stemming from their models and will include appropriate and adequate resources in terms of :

(1) human resources;
(2) financial resources;
(3) access to information and knowledge; and
(4) computational resources.


Resource Allocation for Risk Management
Human Resources
Skilled personnel for tasks
Financial Resources
Funds for operations and investments
Systemic Risks
Addressing risks from models
Information and Knowledge
Access to relevant data
Responsibility
Ensuring effective resource allocation
Computational Resources
Tools for data analysis
1
2
3
4
5
6

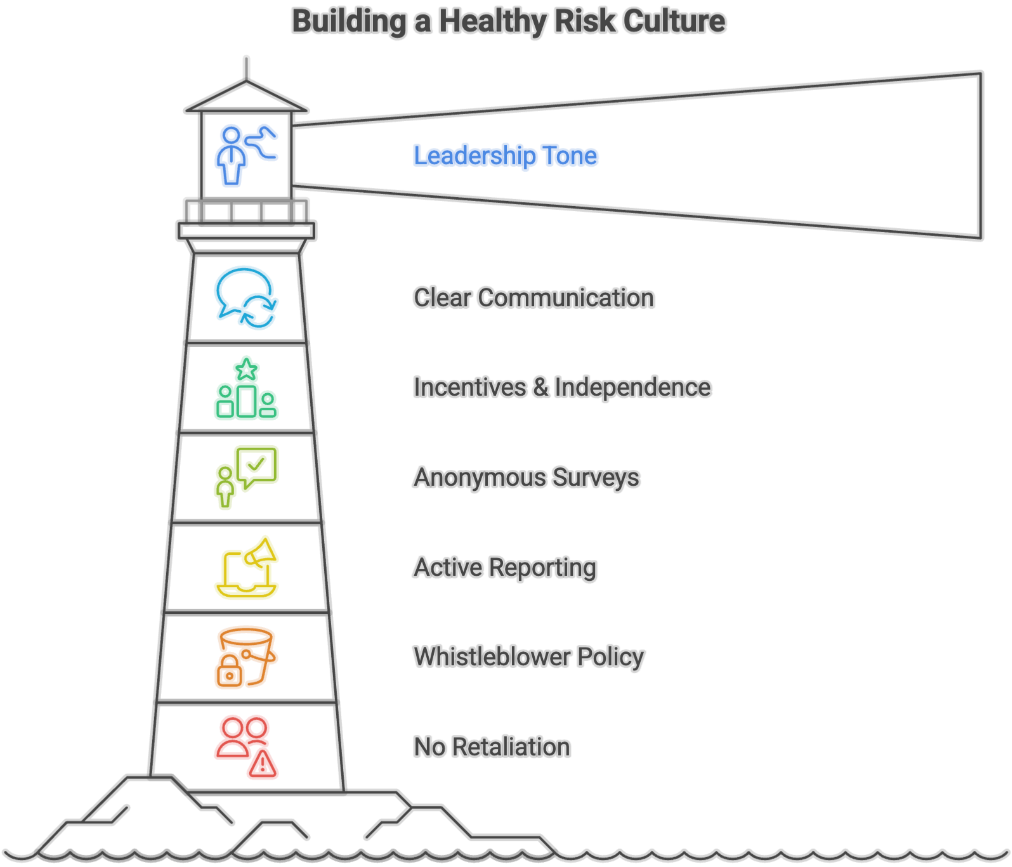
Measure 8.3 Promotion of a healthy risk culture

Providers are required to promote a healthy risk culture and take appropriate measures to ensure that persons who have been assigned responsibilities for managing the systemic risks stemming from their models take a reasoned and balanced approach to systemic risk.

Ensuring a healthy Risk Culture
Companies will therefore have to ensure that they are:

(1) setting the tone for a healthy systemic risk culture from leadership levels[89];
(2) allowing clear communication and challenge of decisions concerning systemic risk[90];
(3) setting incentives and affording sufficient independence of staff involved in systemic risk assessment and mitigation to discourage excessive systemic-risk-taking and encourage an unbiased assessment of the systemic risks stemming from their models;
(4) providing anonymous surveys find that staff are comfortable raising concerns about systemic risks, are aware of channels for doing so, and understand the Signatory's Framework;
(5) ensuring that internal reporting channels are actively used and reports are acted upon appropriately;
(6) annually informing workers of the Signatory's whistleblower protection policy and making such policy readily available to workers such as by publishing it on their website; and/or
(7) not retaliating *in any form*[91] against any person publishing or providing information internally or to competent authorities, in good faith, about systemic risks stemming from models.

[89] by the leadership clearly communicating the Signatory's Framework to staff
[90] This may include being able to challenge the Board decisions without fear of reprisal as well as breaching confidentiality obligations by reporting as whistleblower to the AI Office
[91] including any direct or indirect detrimental action such as termination, demotion, legal action, negative evaluations, or creation of hostile work environments, The caselaw in Europe means that even being moved off a task or having team members make adverse comments following any for of criticism of system risk measure will infringe this provision and may lead to significant dismissal compensation awards.


Building a Healthy Risk Culture
Leadership Tone
Clear Communication
Incentives & Independence
Anonymous Surveys
Active Reporting
Whistleblower Policy
No Retaliation

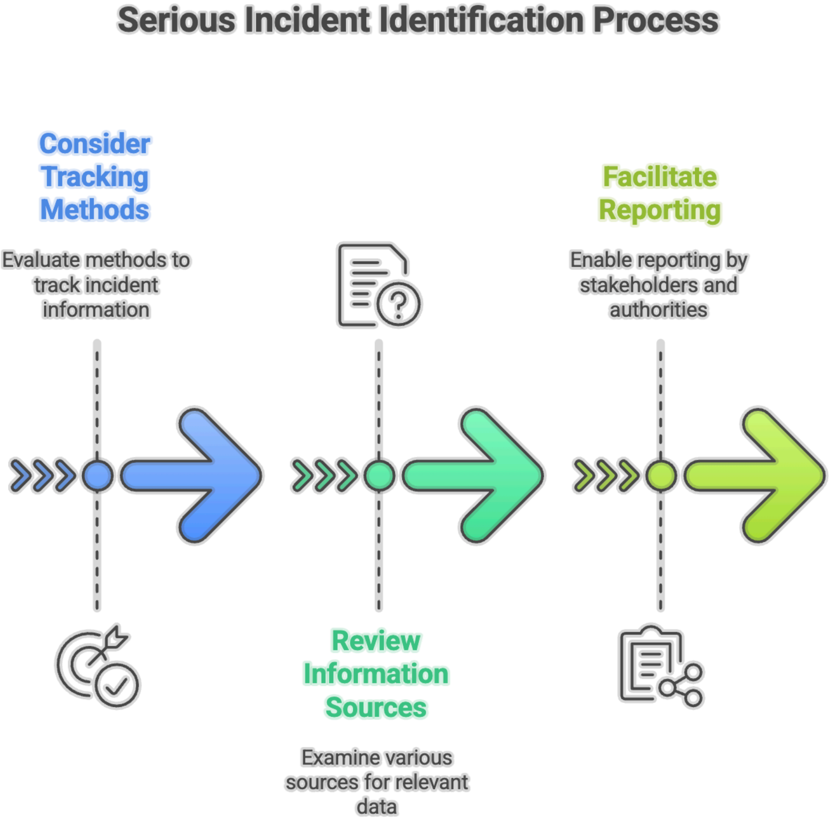
COMMITMENT 9
SERIOUS INCIDENT REPORTING[92]

1. **A commitment 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**.
2. **A commitment to providing resourcing of such processes and measures appropriate for the severity of the serious incident and the degree of involvement of the model.**

Measure 9.1 Methods for serious incident identification

It is necessary to:
1. consider methods to keep track of relevant information about serious incidents;
2. review other sources of information, such as police and media reports, posts on social media, research papers, and incident databases to identify relevant information about serious incidents by downstream modifiers, downstream providers, users, and other third parties; &
3. facilitate the reporting of relevant information about serious incidents by downstream modifiers, downstream providers, users, and other third parties both internally and to the AI Office and, as applicable, national competent authorities including by informing such third parties of available direct reporting channels.

These measure are without prejudice to any of their reporting obligations under Article 73 AI Act.


Serious Incident Identification Process
Consider Tracking Methods
Evaluate methods to track incident information
Review Information Sources
Examine various sources for relevant data
Facilitate Reporting
Enable reporting by stakeholders and authorities

[92] Article 55(1), and recitals 114 and 115 AI Act

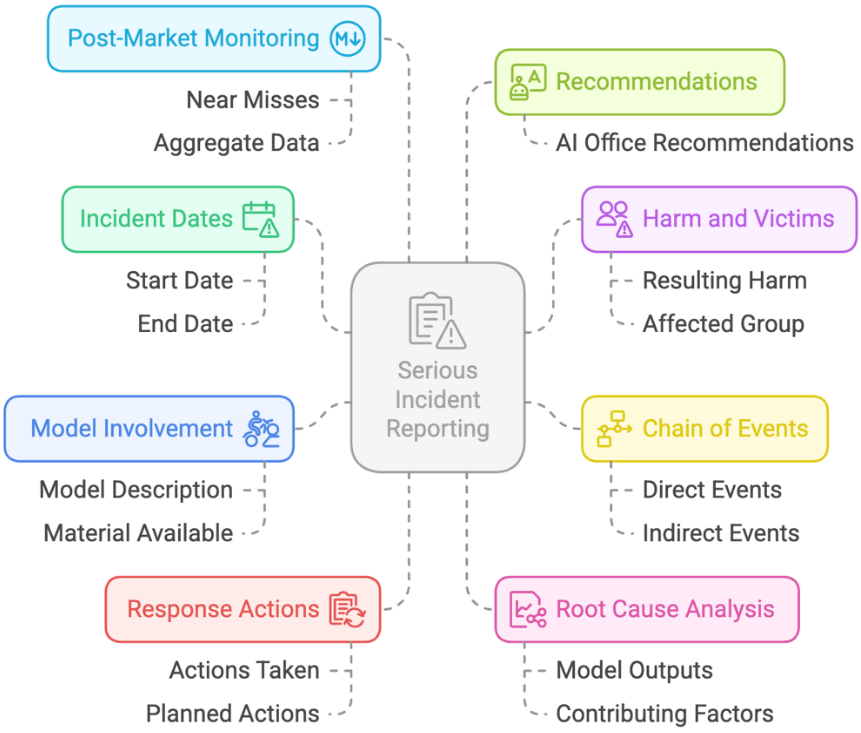
Measure 9.2 Relevant information for serious incident tracking, documentation, and reporting

It is mandatory to keep track of, document, and report to the AI Office and, as applicable, to national competent authorities, at least

(1) the start and end dates of the serious incident, or best approximations thereof if the precise dates are unclear;
(2) the resulting harm and the victim or affected group of the serious incident;
(3) the chain of events that (directly or indirectly) led to the serious incident;
(4) the model involved in the serious incident;
(5) a description of material available setting out the model's involvement in the serious incident;
(6) what, if anything, has been done (or is intended to be done) in response to the serious incident;
(7) what, if any, recommendation is made to the AI Office[93] to do in response to the serious incident;
(8) a root cause analysis with a description of the model's outputs that (directly or indirectly) led to the serious incident and the factors that contributed to their generation, including the inputs used and any failures or circumventions of systemic risk mitigations; and
(9) any patterns detected during post-market monitoring (pursuant to Measure 3.5) that can reasonably be assumed to be connected to the serious incident, such as individual or aggregate data on near misses.

This reporting may be redacted to the extent necessary to comply with other Union law applicable to such information, including confidentiality obligations and trade secrets.

Serious Incident Reporting Requirements



It will be necessary to fully and diligently investigate the causes and effects of serious incidents, including

[93] and, as applicable, national competent authorities

the information within the preceding list, with a view to informing systemic risk assessment.

Where investigations are ongoing so that certain relevant information from the preceding list is not yet available, this must be recorded in the serious incident reports and the report updated when this information is available.

The level of detail in serious incident reports will be appropriate for the severity of the incident.

Measure 9.3 Reporting timelines

The initial report to the AI Office[94] must provide the information listed under Relevant Information above, save in exceptional circumstances, according to the following timetable (determined by when there was awareness of the involvement of the model in the incident) if the model (directly or indirectly) led to:

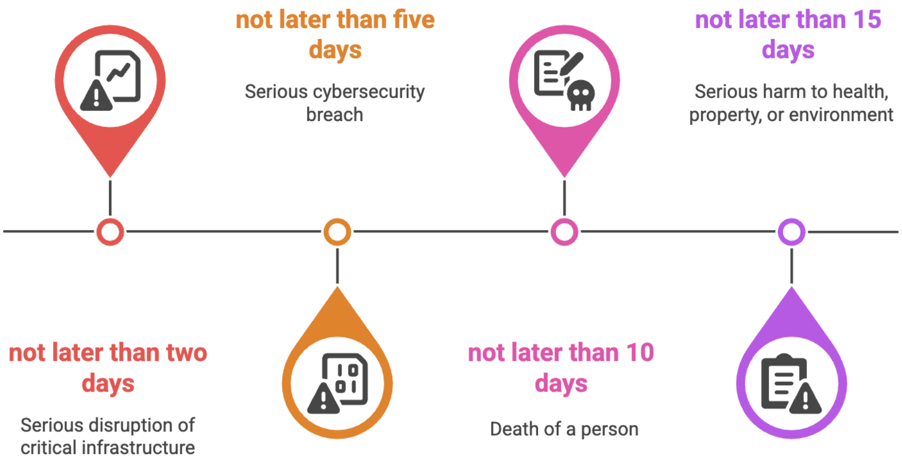
(1) **not later than two days** if a serious and irreversible disruption of the management or operation of critical infrastructure occurred or if it is established or suspected with reasonable likelihood that there was a relevant causal relationship between their model and the disruption;

(2) **not later than five days** if a serious cybersecurity breach[95] occurs or if it is established or suspected with reasonable likelihood such a causal relationship between their model and the breach arose;

(3) **not later than 10 days** if a death of a person is caused or if it is established or suspected with reasonable likelihood such a causal relationship between their model and the death occurred; and

(4) **not later than 15 days if** serious harm to a person's health (mental and/or physical) occurred or there was an infringement of obligations under Union law intended to protect fundamental rights, and/or serious harm to property or the environment, or if it is established or suspected with reasonable likelihood that such a causal relationship between their model and the harms or infringements arose.

Reporting Timelines for AI-Related Incidents


not later than five days
Serious cybersecurity breach
not later than 15 days
Serious harm to health, property, or environment
not later than two days
Serious disruption of critical infrastructure
not later than 10 days
Death of a person

For unresolved serious incidents, the reporting timetables still arise but the reporting in the interim

[94] and, as applicable, to national competent authorities
[95] including the (self-)exfiltration of model weights and cyberattacks

initial report must be updated as soon as reasonably available with updates at least every four weeks after the initial report until the report is final. The final report, covering all the information required must be submitted to the AI Office[96] not later than 60 days after the serious incident has been resolved.

If multiple similar serious incidents occur within the reporting timelines, Signatories may include them in the report(s) of the first serious incident, although this will not change the timetables for reporting for the first serious incident.

Measure 9.4 Retention period

It will be necessary to keep documentation of all relevant information gathered under this Commitment for at least five years from the date of the documentation (i.e. the final report) or the date of the serious incident, whichever is later[97].

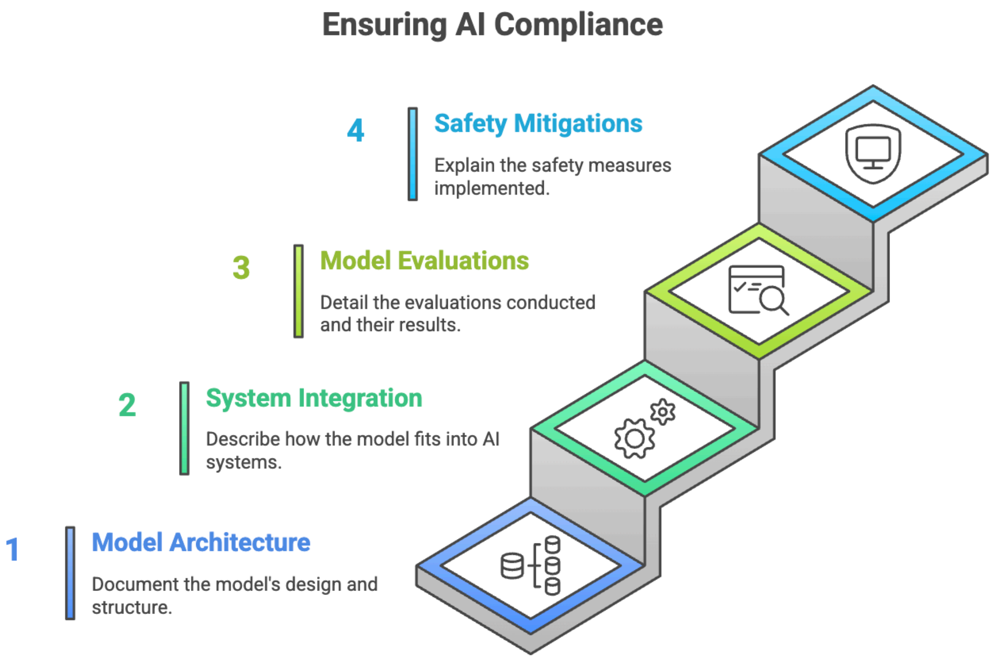
[96] and, as applicable, to national competent authorities

[97] This is without prejudice to Union law applicable to such information, which may mandate longer periods of retention..

COMMITMENT 10
ADDITIONAL DOCUMENTATION AND TRANSPARENCY[98]

1. **A Commitment to documenting the model and publishing a summarised versions of the model Framework and Model Reports as necessary.**

Measure 10.1 Additional documentation


Ensuring AI Compliance
4
Safety Mitigations
Explain the safety measures implemented.
3
Model Evaluations
Detail the evaluations conducted and their results.
2
System Integration
Describe how the model fits into AI systems.
1
Model Architecture
Document the model's design and structure.

This is a commitment to draw up and keep up-to-date the following information for the purpose of providing it to the AI Office upon request:

(1) a detailed description of the model's architecture;

(2) a detailed description of how the model is integrated into AI systems, explaining how software components build or feed into each other and integrate into the overall processing, insofar as the reporting party is aware of such information;

(3) a detailed description of the model evaluations conducted under the Codes, including their results and strategies; and

(4) a detailed description of the safety mitigations implemented.

[98] Articles 53(1)(a) and 55(1) AI Act

Documentation must be retained at least 10 years after the model has been placed on the market.

The following information, to the extent it is not already provided will be retained and provided to the AI Office upon request:

(1) the processes, measures, and key decisions that form part of systemic risk assessments and mitigations;
and

(2) justifications for choices of a particular best practice, state-of-the-art, or other more innovative processes or measures relied upon as adherence to this Code.

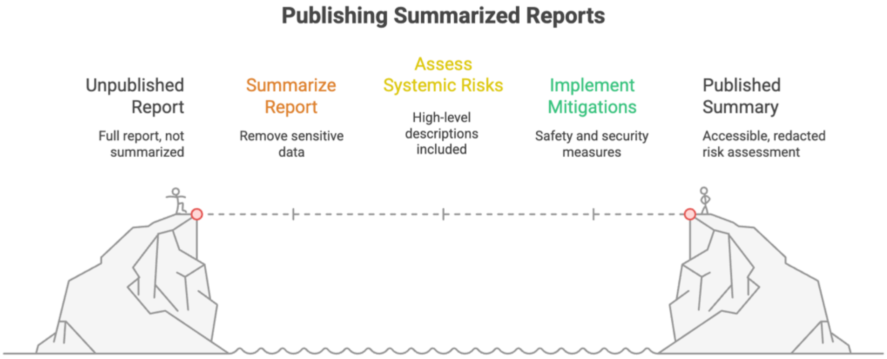
Measure 10.2 Public transparency

It may be necessary to publish[99] a summarised version of the Framework and Model Report(s), (and updates thereof) in order to allow assessment and/or mitigate of systemic risks; however these publications may have data removed or redacted in otder not undermine the effectiveness of safety and/or security mitigations and/or 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.


Publishing Summarized Reports
Unpublished Report
Full report, not summarized
Summarize Report
Remove sensitive data
Assess Systemic Risks
High-level descriptions included
Implement Mitigations
Safety and security measures
Published Summary
Accessible, redacted risk assessment

[99] Via the website or other standard procedures used for publishing information relating to the models

THIS IS THE GLOSSARY AS PUBLISHED IN THE SAFETY AND SECURITY CODE OF PRACTICE

Terms defined in Article 3 AI Act, shall apply in relation to definitions and these shall prevail, otherwise and complementing this, the following terms shall apply. Unless otherwise stated, all grammatical variations of the terms defined in this Glossary shall be deemed to be covered by the relevant definition.

Term	Definition
‘appropriate’	suitable and necessary to achieve the intended purpose of systemic risk assessment and/or mitigation, whether through best practices, the state of the art, or other more innovative processes, measures, methodologies, methods, or techniques that go beyond the state of the art.
‘best practice’	accepted amongst providers of general-purpose AI models with systemic risk as the processes, measures, methodologies, methods, and techniques that best assess and mitigate systemic risks at any given point in time.
‘confirmed’	a Framework or Model Report, or an update thereof, that has received required approvals under the applicable governance procedures.
‘deception’	model behaviours that systematically produce false beliefs in others, including model behaviours to achieve goals that involve evading oversight, such as a model’s detecting that it is being evaluated and under-performing or otherwise undermining oversight.
‘external validity’	an aspect of high scientific and technical rigour (see definition below) that ensures model evaluations are suitably calibrated for results to be used as a proxy for model behaviour outside the evaluation environment. Demonstrating external validity will differ for different systemic risks and model evaluation methods, but may be shown by, e.g. documenting the model evaluation environment, the ways in which it diverges from the real-world context, and the diversity of the model evaluation environment.
‘high scientific and technical rigour’	the quality standard for model evaluations, such that model evaluations with high scientific and technical rigour have internal validity (see definition below) and external validity (see definition above), as well as being reproducible (see definition below). See further Appendix 3.2.
‘including’	introduces a non-exhaustive set that is to be understood as the minimum required by the term referred to and is indicative of further items of the set.
‘independent external’	a natural or legal person that has no financial, operational, or management dependence on the Signatory or any of its subsidiaries or associates, and is otherwise free from the Signatory’s control in reaching conclusions and/or making recommendations, including through contractual safeguards and suitable conflict of interest policies.

‘insider threats’	hostile operations by humans, AI models, and/or AI systems (e.g. senior management, a senior member of the organisation’s research team, other disgruntled employees, perpetrators of industrial espionage operations that have infiltrated their target, and/or model self-exfiltration) with access to sensitive organisational resources, and/or accidental model leakage.
‘internal validity’	an aspect of high scientific and technical rigour (see definition above) that ensures model evaluation results are as accurate as scientifically possible in the evaluation setting and are free from methodological shortcomings that could undermine the results. Demonstrating internal validity will differ for different systemic risks and model evaluation methods, but may be shown by, e.g.: large enough sample sizes; measuring statistical significance and statistical power; disclosure of environmental parameters used; controlling for confounding variables and mitigating spurious correlation; preventing use of test data in training (e.g. using train-test splits and respecting canary strings); re-running model evaluations multiple times under different conditions and in different environments, including varying individual parts of model evaluations (e.g. the strength of prompts and safety and security mitigations); detailed inspection of trajectories and other outputs; avoiding potential labelling bias in model evaluations, particularly model evaluations involving human annotators (e.g. through blinding or reporting inter-annotator agreement); using transparency-increasing techniques (e.g. reasoning traces in evaluations that are representative of the model’s “inner workings” and legible by evaluators); using techniques to measure and/or reduce the model’s capability to evade oversight; and/or disclosing the methods for creating and managing new model evaluations to ensure their integrity.
‘management body’	a corporate organ appointed pursuant to national law and empowered to perform: (1) an executive function by (a) setting the organisation’s strategy, objectives, and overall direction, and (b) conducting day-to-day management of the organisation; and (2) a supervisory function by overseeing and monitoring executive decision-making. Depending on the relevant national law, the executive and supervisory functions may be performed by different personnel within the one management body or they may be performed by distinct parts of the management body.
‘model’	a general-purpose AI model with systemic risk. There may be many different versions of the same model, such as versions fine-tuned for different purposes, versions with access to different tools, and/or versions with different safety and/or security mitigations. All references to ‘model’ in this Chapter refer to the relevant model version(s), as the context requires. Generally, in the context of systemic risk assessment and mitigation, all references to ‘model’ refer to all model versions that, in aggregate, constitute the systemic

	risk(s) stemming from the model, including all model versions that: (1) are the most advanced; (2) correspond to point (1) and have limited or no safety and/or security mitigations for systemic risk implemented; and/or (3) are used widely. In the context of comparisons between different 'models' (such as in Measure 3.5 and Appendix 3.5, in conjunction with Appendix 2, and Commitment 6), all references to 'model' refer to a single model version. If the term 'AI' precedes the term 'model', this term exceptionally does not only refer to general-purpose AI models with systemic risk but also includes all models other than general-purpose AI models with systemic risk.
'model elicitation'	technical work to systematically enhance a model's capabilities, propensities, affordances, and/or effects, thereby facilitating an accurate measurement of the full range of its capabilities, propensities, affordances, and/or effects that can likely be attained.
'model evaluation'	a systemic risk assessment technique that can be used in all stages of systemic risk assessment (as defined below).
'model-independent information'	information, including data and research, that is not tied to a specific model, but can inform systemic risk assessment and mitigation across several models. See further Measure 3.1.
'near miss'	a situation in which a serious incident could have, but ultimately did not, materialise.
'non-state external threats'	hostile operations conducted by non-state actors that: (1) are roughly comparable to ten experienced, professional individuals in cybersecurity; (2) spend several months with a total budget of up to EUR 1 million on the specific operation; and (3) have major pre-existing cyberattack infrastructure but no pre-existing access to the target organisation.
'post-market monitoring'	the monitoring of a model in the time span from when it is placed on the market until the retirement of the model from being made available on the market. See further Measure 3.5.
'process' (noun; in the context of systemic risk management)	a structured set of actions that comprise or result in measures stipulated by this Chapter.
'reproducibility'	an aspect of high scientific and technical rigour (see definition above) that refers to the ability to obtain consistent model evaluation results using the same input data, computational techniques, code, and model evaluation conditions, allowing for other researchers and engineers to validate, reproduce, or improve on model evaluation results.

	Reproducibility may be shown by, e.g.: successful peer reviews and/or reproductions by independent third parties; securely releasing to the AI Office adequate amounts of model evaluation data, model evaluation code, documentation of model evaluation methodology and methods, model evaluation environment and computational environment, and model elicitation techniques; and/or use of publicly available APIs, technical model evaluation standards, and tools.
‘resolved’ (serious incident)	a serious incident of a model for which the Signatory adopted corrective measures to rectify the harm, if possible, and to assess and mitigate systemic risks related to it. ‘Unresolved’ is to be understood accordingly.
‘scaling law’	a systematic relationship between some variable relevant to the development or use of an AI model or AI system, such as size or the amount of time, data, or computational resources used in training or inference, and its performance.
‘(self-)exfiltration of model weights’	access or transfer of weights or associated assets of a model from their secure storage by the model itself and/or an unauthorised actor.
‘similar model’	a general-purpose AI model with or without systemic risk, assumed to have materially similar capabilities, propensities, and affordances based on public and/or private information available to the Signatory, including “safe reference models” (pursuant to Appendix 2.1) and “similarly safe or safer models” (pursuant to Appendix 2.2).
‘state of the art’	the forefront of relevant research, governance, and technology that goes beyond best practice.
‘system prompt’	a set of instructions, guidelines, and contextual information provided to a model before a user interaction begins.
‘systemic risk acceptance criteria’	criteria defined in the Framework that Signatories use to decide whether the systemic risks stemming from their models are acceptable. Systemic risk tiers (as defined below) are a type of systemic risk acceptance criteria. See further Measure 4.1.
‘systemic risk assessment’	the overarching term referring to all of systemic risk identification (pursuant to Commitment 2), systemic risk analysis (pursuant to Commitment 3), and systemic risk acceptance determination (pursuant to Commitment 4).
‘systemic risk management’	coordinated processes and measures to direct an organisation with regard to systemic risk, including systemic risk assessment and mitigation.
‘systemic risk mitigations’	comprise safety mitigations (pursuant to Commitment 5), security mitigations (pursuant to Commitment 6), and governance mitigations (pursuant to Commitments 1 and 7 to 10) for systemic risk.
‘systemic risk modelling’	a structured process aimed at specifying pathways through which a systemic risk stemming from a model might materialise; often used interchangeably with the term ‘threat modelling’. This Chapter uses the term ‘risk modelling’ because the term ‘threat modelling’ has a specific meaning in cybersecurity. See further Measure 3.3.

‘systemic risk scenario’	a scenario in which a systemic risk stemming from a model might materialise. See further Measure 2.2.
‘systemic risk source’	a factor which alone or in combination with other factors might give rise to systemic risk. See further Appendix 1.3.
‘systemic risk tiers’	tiers defined in the Framework that corresponds to a certain level of systemic risk stemming from a model. Systemic risk tiers are a type of systemic risk acceptance criteria. See further Measure 4.1.
‘use’ (of a model)	use of the model by the Signatory or other actors.

THESE ARE THE APPENDICES AS PUBLISHED IN THE SAFETY AND SECURITY CODE OF PRACTICE

APPENDIX 1 SYSTEMIC RISKS AND OTHER CONSIDERATIONS

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 1.1 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 1.2 NATURE OF SYSTEMIC RISKS

The considerations below concerning the nature of systemic risks inform systemic risk identification (pursuant to Commitment 2). The considerations distinguish between essential characteristics of the nature of systemic risks (Appendix 1.2.1) and contributing characteristics (Appendix 1.2.2).

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.

APPENDIX 1.2.2 CONTRIBUTING CHARACTERISTICS

(1) **Capability-dependent:** The risk increases with model capabilities or may emerge at the frontier of model capabilities.
(2) **Reach-dependent:** The risk increases with model reach.
(3) **High velocity:** The risk can materialise rapidly, potentially outpacing mitigations.
(4) **Compounding or cascading:** The risk can trigger other systemic risks or chain reactions.
(5) **Difficult or impossible to reverse:** Once materialised, the risk creates persistent changes that require extraordinary effort, resources, or time to remediate, or are permanently irreversible.
(6) **Asymmetric impact:** A small number of actors or events can trigger the materialisation of the risk, causing disproportionate impact relative to the number of actors or events.

APPENDIX 1.3 SOURCES OF SYSTEMIC RISKS

The following model capabilities, model propensities, model affordances, and contextual factors are treated as non-exhaustive, potential systemic risk sources for the purpose of systemic risk identification (pursuant to Commitment 2).

APPENDIX 1.3.1 MODEL CAPABILITIES

Model capabilities include:

(1) offensive cyber capabilities;
(2) Chemical, Biological, Radiological, and Nuclear (CBRN) capabilities, and other such weapon acquisition or proliferation capabilities;
(3) capabilities that could cause the persistent and serious infringement of fundamental rights;
(4) capabilities to manipulate, persuade, or deceive;
(5) capabilities to operate autonomously;
(6) capabilities to adaptively learn new tasks;
(7) capabilities of long-horizon planning, forecasting, or strategising;
(8) capabilities of self-reasoning (e.g. a model's ability to reason about itself, its implementation, or environment, its ability to know if it is being evaluated);
(9) capabilities to evade human oversight;
(10) capabilities to self-replicate, self-improve, or modify its own implementation environment;
(11) capabilities to automate AI research and development;
(12) capabilities to process multiple modalities (e.g. text, images, audio, video, and further modalities);
(13) capabilities to use tools, including "computer use" (e.g. interacting with hardware or software that is not part of the model itself, application interfaces, and user interfaces); and
(14) capabilities to control physical systems.

APPENDIX 1.3.2 MODEL PROPENSITIES

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

(1) misalignment with human intent;
(2) misalignment with human values (e.g. disregard for fundamental rights);
(3) tendency to deploy capabilities in harmful ways (e.g. to manipulate or deceive);
(4) tendency to "hallucinate", to produce misinformation, or to obscure sources of information;
(5) discriminatory bias;
(6) lack of performance reliability;
(7) lawlessness, i.e. acting without reasonable regard to legal duties that would be imposed on similarly situated persons, or without reasonable regard to the legally protected interests of affected persons;
(8) "goal-pursuing", harmful resistance to goal modification, or "power-seeking";
(9) "colluding" with other AI models/systems; and
(10) mis-coordination or conflict with other AI models/systems.

APPENDIX 1.3.3 MODEL AFFORDANCES AND OTHER SYSTEMIC RISK SOURCES

Model affordances and other systemic risk sources, encompassing model configurations, model properties, and the context in which the model is made available on the market, include:

(1) access to tools (including other AI models/systems), computational power (e.g. allowing a model to increase its speed of operations), or physical systems including critical infrastructure;
(2) scalability (e.g. enabling high-volume data processing, rapid inference, or parallelisation);
(3) release and distribution strategies;
(4) level of human oversight (e.g. degree of model autonomy);
(5) vulnerability to adversarial removal of guardrails;
(6) vulnerability to model exfiltration (e.g. model leakage/theft);
(7) lack of appropriate infrastructure security;
(8) number of business users and number of end-users of the model, including the number of end- users using an AI system in which the model is integrated;
(9) offence-defence balance, including the potential number, capacity, and motivation of malicious actors to misuse the model;
(10) vulnerability of the specific environment potentially affected by the model (e.g. social environment, ecological environment);
(11) lack of appropriate model explainability or transparency;
(12) interactions with other AI models and/or AI systems; and
(13) inappropriate use of the model (e.g. using the model for applications that do not match its capabilities or propensities).

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 2 SIMILARLY SAFE OR SAFER MODELS

APPENDIX 2.1 SAFE REFERENCE MODELS

A model may be considered a safe reference model with regard to a systemic risk if:

(1) the model has: (a) been placed on the market before the publication of this Chapter; or (b) completed the full systemic risk assessment and mitigation process (pursuant to Measure 1.2, third paragraph), including that the systemic risks stemming from the model have been determined to be acceptable (pursuant to Commitment 4), and the AI Office has received its Model Report (pursuant to Commitment 7);
(2) the Signatory has sufficient visibility into the model's characteristics such as relevant architectural details, capabilities, propensities, affordances, and safety mitigations. Such visibility is assumed for all models developed by the Signatory itself and for all models for which the Signatory has access to all information that would be necessary for the Signatory to complete the full systemic risk assessment and mitigation process (pursuant to Measure 1.2, third paragraph), including the model parameters; and
(3) there are no other reasonable grounds to believe that the systemic risks stemming from the model are not acceptable.

APPENDIX 2.2 SIMILARLY SAFE OR SAFER MODELS

A model may be considered a similarly safe or safer model with regard to a systemic risk if:

(1) Signatories do not reasonably foresee any materially different systemic risk scenario (pursuant to Measure 2.2) regarding the systemic risk for the model compared to the safe reference model after conducting systemic risk identification (pursuant to Commitment 2);
(2) the scores of the model on relevant at least state-of-the-art, light-weight benchmarks are all lower than or equal to (within a negligible margin of error) the scores of the safe reference model. Minor increases in capabilities compared to the safe reference model that result in no material increase in the systemic risk may be disregarded. Such benchmarks must have been run pursuant to Measure
3.2; and
(3) there are no known differences in the model's characteristics such as relevant architectural details, capabilities, propensities, affordances, and safety mitigations compared to the safe reference model that could be reasonably foreseen to result in a material increase in the systemic risk, and there are no other reasonable grounds to believe that the systemic risks stemming from the model are materially increased compared to the safe reference model.

In making their assessment of points (2) and (3) in the preceding paragraph and Appendix 2.1, point (2), Signatories will appropriately take into account the uncertainty that may stem from, e.g. a lack of information about the reference model and measurement errors, by incorporating a sufficiently wide safety margin.

In the event that a model previously considered to be a safe reference model by the Signatory for treating another model as a similarly safe or safer model subsequently loses this status as safe reference model, the Signatory will within six months:
(1) identify another safe reference model in relation to which the model may be considered a similarly safe or safer model; or
(2) treat the other model as subject to all Commitments and Measures of this Chapter if previously adherence had relied on exemptions and/or reductions by virtue of its similarly safe or safer status, including completing all previously exempted and/or reduced parts of the full systemic risk assessment and mitigation process (pursuant to Measure 1.2, third paragraph).

APPENDIX 3 MODEL EVALUATIONS

The following specifies the model evaluations required by Measure 3.2 during the full systemic risk assessment and mitigation process (pursuant to Measure 1.2, third paragraph).

APPENDIX 3.1 RIGOROUS MODEL EVALUATIONS

Signatories will ensure that the model evaluations are conducted with high scientific and technical rigour, ensuring:
(1) internal validity;
(2) external validity; and
(3) reproducibility.

APPENDIX 3.2 MODEL ELICITATION

Signatories will ensure that the model evaluations are conducted with at least a state-of-the-art level of model elicitation that elicits the model's capabilities, propensities, affordances, and/or effects, by using at least state-of-the-art techniques that:

(1) minimise the risk of under-elicitation; and
(2) minimise the risk of model deception during model evaluations (e.g. sandbagging);

such as by adapting test-time compute, rate limits, scaffolding, and tools, and conducting fine-tuning and prompt engineering.

For this, Signatories will at least:

(1) match the model elicitation capabilities of misuse actors relevant to the systemic risk scenario (pursuant to Measure 2.2); and
(2) match the expected use context (e.g. equivalent scaffolding and/or tool access) of the model, as informed by integrations into AI systems that are:
    (a) planned or considered for the model; and/or
    (b) currently used for similar models, if such integrations are known to the Signatory and the Signatory cannot exclude a similar use of their model.

APPENDIX 3.3 ASSESSING THE EFFECTIVENESS OF MITIGATIONS

Signatories will ensure that the model evaluations assess the effectiveness of their safety mitigations at a breadth and depth appropriate for the extent to which systemic risk acceptance determination depends on the effectiveness of specific mitigations, including under adversarial pressure (e.g. fine-tuning attacks or jailbreaking). To this end, Signatories will use at least state-of-the-art techniques, taking into account:

(1) the extent to which their mitigations work as planned;
(2) the extent to which their mitigations are or have been circumvented, deactivated, or subverted; and
(3) the probability that the effectiveness of their mitigations will change in the future.

APPENDIX 3.4 QUALIFIED MODEL EVALUATION TEAMS AND ADEQUATE RESOURCES

Signatories will ensure that the teams responsible for conducting the model evaluations combine technical expertise with relevant domain knowledge of the systemic risk to enable a holistic and multi-disciplinary understanding. Indicative qualifications for such technical expertise and/or relevant domain knowledge are:

(1) a PhD, peer-reviewed and recognised publications, or equivalent research or engineering experience, relevant to the systemic risk;
(2) having designed or developed a published, and peer-reviewed or widely used, model evaluation method for the systemic risk; or
(3) three years of work-experience in a field directly relevant to the systemic risk or, if that field is nascent, equivalent experience from studying in the field or working in a field with directly transferable knowledge.

Model evaluation teams will be provided with:

(1) adequate access to the model to conduct the model evaluations pursuant to this Appendix 3, including, as appropriate, access to model activations, gradients, logits (or other forms of raw model outputs), chains-of-thought, and/or other technical details, and access to the model version(s) with the fewest safety mitigations implemented (such as a helpful-only model version, if it exists). Regarding the adequacy of heightened model access for model evaluation teams, Signatories will take into account the potential risks to model security that this can entail and implement appropriate security measures for the evaluations;

(2) information, including model specifications (including the system prompt), relevant training data, test sets, and past model evaluation results, as appropriate for: (a) the systemic risk; and (b) the model evaluation method;

(3) time to competently design and/or adapt, debug, execute, and analyse the model evaluations pursuant to this Appendix 3, as appropriate for: (a) the systemic risk; and (b) the model evaluation method and its novelty. For example, a period of at least 20 business days is appropriate for most systemic risks and model evaluation methods; and

(4) (a) adequate compute budgets, including to allow for sufficiently long model evaluation runs, parallel execution, and re-runs; (b) adequate staffing; and (c) adequate engineering budgets and support, including to inspect model evaluation results to identify and fix software bugs or model refusals which might lead to artificially lowered capability estimates. With respect to point (b), if Signatories engage independent external evaluators, they may rely on the latter's assurances as to whether their staffing is adequate.

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.

Signatories that are SMEs or SMCs may contact the AI Office, which may provide support or resources to facilitate adherence to this Appendix 3.5.

APPENDIX 4 SECURITY MITIGATION OBJECTIVES AND MEASURES

The following specifies the security mitigation objectives and measures (pursuant to Measure 6.2) to be implemented in order to meet the Security Goal.

APPENDIX 4.1 GENERAL SECURITY MITIGATIONS

Signatories will implement general security mitigations that achieve the following mitigation objectives:

(1) prevention of unauthorised network access, through (a) strong identity and access management practices, including restrictions on device and account sharing, multi-factor authentication, strong password enforcement, strong access management tools, 802.1x authentication, zero trust architecture, protection of wireless networks to the same standard as wired networks, and the separation of any guest networks from the work network;

(2) reduction of the risk of social engineering, through (a) email filtering for suspicious attachments, links, and other phishing attempts;

(3) reduction of the risk of malware infection and malicious use of portable devices, through (a) policies regarding the use of removable media; and

(4) reduction of the risk of vulnerability exploitation and malicious code execution, through (a) regular software updates and patch management.

APPENDIX 4.2 PROTECTION OF UNRELEASED MODEL PARAMETERS

Signatories will protect unreleased model parameters by implementing security mitigations that achieve the following mitigation objectives:

(1) accountability over all copies of stored model parameters across all devices and locations, through (a) a secure internal registry of all devices and locations where model parameters are stored;

(2) prevention of unauthorised copying of model parameters to unmanaged devices, through (a) access management on all devices storing model parameters, with alerts in case of copying to unmanaged devices;

(3) prevention of unauthorised access to model parameters during transport and at rest, through (a) ensuring model parameters are always encrypted during transportation and storage as appropriate, including encryption with at least 256-bit security and with encryption keys stored securely on a Trusted Platform Module (TPM);

(4) prevention of unauthorised access to model parameters during temporary storage, through (a) ensuring model parameters are only decrypted for legitimate use to non-persistent memory;

(5) prevention of unauthorised access to model parameters during use, through (a) implementing confidential computing as appropriate, using hardware-based, and attested trusted execution environments; and

(6) prevention of unauthorised physical access to systems hosting model parameters, through (a) restricting physical access to data centres and other sensitive working environments to required personnel only, along with regular inspections of such sites for unauthorised personnel or devices.

APPENDIX 4.3 HARDENING INTERFACE-ACCESS TO UNRELEASED MODEL PARAMETERS

Signatories will harden interface-access to unreleased model parameters while in use, by implementing security mitigations that achieve the following mitigation objectives:

(1) prevention of unnecessary interface-access to model parameters, through (a) explicitly authorising only required software and persons for access to model parameters, enforced through multi-factor authentication mechanisms, and checked on a regular basis of at least every six months;

(2) reduction of the risk of vulnerability exploitation or data leakage, through (a) thorough review of any software interfaces with access to model parameters by a security team to identify vulnerabilities or data leakage, and/or automated security reviews of any software interface code

at least to the same standard as the highest level of automated security review used for other sensitive code;

(3) reduction of the risk of model parameter exfiltration, through (a) hardening interfaces with access to model parameters, using methods such as output rate limiting; and

(4) reduction of the risk of insider threats or compromised accounts, through (a) limiting the number of people who have non-hardened interface-access to model parameters.

APPENDIX 4.4 INSIDER THREATS

Signatories will protect against insider threats, including in the form of (self-)exfiltration or sabotage carried out by models, by implementing security mitigations that achieve the following mitigation objectives:

(1) protection of model parameters from insider threats attempting to gain work-related access with the Signatory, through (a) background checks on employees and contractors that have or might reasonably obtain read or write access to unreleased model parameters or systems that manage the access to such parameters;

(2) awareness of the risk of insider threats, through (a) the provision of training on recognising and reporting insider threats;

(3) reduction of the risk of model self-exfiltration, through (a) sandboxes around models, such as virtual machines and code execution isolation; and

(4) reduction of the risk of sabotage to model training and use, through (a) checking training data for indications of tampering.

APPENDIX 4.5 SECURITY ASSURANCE

Signatories will obtain assurance that their security mitigations meet the Security Goal by implementing additional security mitigations that achieve the following mitigation objectives:

(1) independent external validation of security mitigation effectiveness if internal expertise is inadequate, through (a) regular independent external security reviews as appropriate to mitigate systemic risks;

(2) validation of network and physical access management and security gap identification, through (a) frequent red-teaming as appropriate to mitigate systemic risks;

(3) validation of network software integrity, through (a) competitive bug bounty programs to encourage public participation in security testing of public-facing endpoints as appropriate to mitigate systemic risks;

(4) validation of insider threat security mitigations, through (a) periodic personnel integrity testing;

(5) facilitation of reporting of security issues, through (a) secure communication channels for third parties to report security issues;

(6) detection of suspicious or malicious activity, through (a) installation of Endpoint Detection and Response (“EDR”) and/or Intrusion Detection System (IDS) tools on all networks and devices; and

(7) timely and effective response to malicious activity, through (a) the use of a security team to monitor for EDR alerts and conduct security incident handling, response, and recovery for security breaches in a timely and effective manner.