# AI Safeguard:

# Navigating Compliance and Risk in the Era of the EU AI Act

Heramb Podar

IIT Roorkee

**Organized by**
Charlotte Siegmann, Esben Kran

## Abstract

The EU AI Act heralds a transformative era in AI governance, mandating rigorous quality management and extensive technical documentation from AI system providers. Yet, the challenge of crafting a comprehensive risk management framework that not only systematically pinpoints and assesses risks but also seamlessly aligns with the Act's mandates looms large.
Our proposed framework addresses the critical need for an effective risk management strategy that aligns with the EU AI Act. It offers providers a clear, practical guide for managing risks in AI systems, ensuring compliance in an increasingly regulated AI landscape. This guidance is designed to be a key tool in achieving responsible AI deployment.

*Keywords: EU AI Act , Risk management, monitoring*

## 1. Scope

For the purposes of this implementation guidance, we have limited our scope of concern to proposing a Risk Management Framework for high-risk AI systems as defined by Title III and Annex III of the EU AI Act. Definitions for terms can be found in Appendix 1, which primarily sources definitions from the EU AI Act and relevant ISO standards.

## 3. Introduction

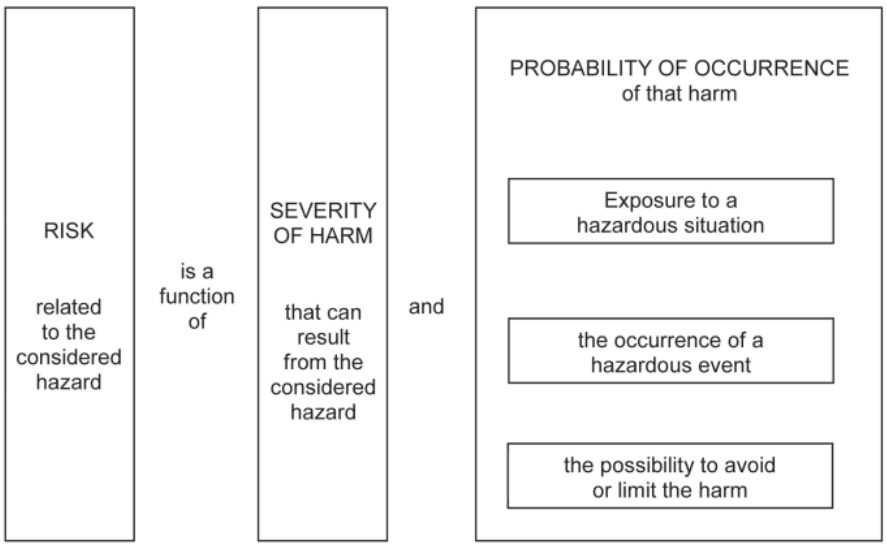
The purpose of this implementation guidance is to serve as a handbook for providers of high-risk AI systems to make compliant risk management frameworks pursuant to Articles 9 and Annexure-VII of the EU AI Act. The Risk Management Framework outlined below covers the chain of risk management starting from risk identification to the setting up of a continuous improvement for compliance management system.

The three core problems at the heart of the regulatory challenge of frontier AI models are the Deployment Safety Problem, Proliferation Problem and the Unexpected Capabilities Problem. We believe that this risk management framework also addresses parts of the Deployment Safety Problem and can be extended as a layer of defence against the other two.

Providers of high-risk systems must obtain a CE stamp for their systems to get certified market access to the European Economic Area (EEA), enhance consumer trust and achieve regulatory compliance. To this end, providers of high-risk AI systems must submit a quality management system for conformity assessment. This quality management system requires a risk management framework strategy in order to strategically mitigate any risks pertaining to safety, and healthcare.

## 4. Implementation Guidance

1. Risk Identification:



*Risk as defined by ISO/IEC 51*

**Objective:** To systematically identify potential risks in AI systems, including both evident and less apparent hazards.

**Process:**

Listing Known Risks:

- Catalog common risks like algorithmic errors, unauthorized data access, and biased decision-making.

Brainstorming with Experts:

- Conduct sessions with AI experts and various stakeholders to delve into less obvious risks.

Real-World Examples:

- Apply practical scenarios to contextualize risks, such as algorithmic bias in AI-driven credit scoring systems potentially leading to unfair scoring decisions.

Defining Risk Criteria:

- Proper risk criteria must be defined on the basis of training, the capabilities of the AI system and the risk appetite of the provider.

| Considerations for defining risk criteria as provided in ISO 31000:2018, 6.3.4 | Additional considerations in the context of the development and use of AI systems |
|---|---|
| — The nature and type of uncertainties that can affect outcomes and objectives (both tangible and intangible); | — Organizations should take reasonable steps to understand uncertainty in all parts of the AI system, including the utilized data, software, mathematical models, physical extension, and human-in-the-loop aspects of the system (such as any related human activity during data collection and labelling). |
| — How consequences (both positive and negative) and likelihood will be defined and measured; | |
| — Time-related factors; | — No specific guidance beyond ISO 31000:2018 |
| — Consistency in the use of measurements; | — Organizations should be aware that AI is a fast-moving technology domain. Measurement methods should be consistently evaluated according to their effectiveness and appropriateness for the AI systems in use. |
| — How the level of risk is to be determined; | — Organizations should establish a consistent approach to determine the risk level. The approach should reflect the potential impact of AI systems regarding different AI-related objectives (see Annex A). |
| — How combinations and sequences of multiple risks will be taken into account; | — No specific guidance beyond ISO 31000:2018 |
| — The organization's capacity. | — The organization's AI capacity, knowledge level and ability to mitigate realized AI risks should be considered when deciding its AI risk appetite. |

*ISO 31000 Risk criteria example*

2. Risk Analysis:

**Objective:** To evaluate the likelihood and impact of identified risks in the laboratory environment using a blend of qualitative and quantitative methods.

**Process:**

Qualitative Analysis:

- Conduct expert panel discussions to assess the probability and impact of each identified risk based on experience and knowledge.
- Use tools like risk matrices to categorize and prioritize risks.

Quantitative Analysis:

- Employ statistical methods to estimate the probability of risk occurrence and its potential impact.

Example: For a data breach risk, calculate potential financial loss based on past incidents and industry benchmarks. Assess reputational damage by analyzing customer feedback and case studies from similar incidents.

Documenting the Analysis:

- Create detailed records of both qualitative and quantitative analyses, including methodologies used, assumptions made, and results.

Review and Update:

- Regularly review and update the risk analysis to reflect new data, changes in the lab environment, or emerging risks.
- By meticulously analyzing risks through these methods, laboratories can better understand their risk profile and prepare appropriate mitigation strategies.

3. Risk Evaluation:

**Objective:** Methodically prioritize risks based on their impact and likelihood, aligning with the laboratory's defined risk appetite, focusing first on those that could significantly impact safety, compliance, or critical system functionality.

**Process:**

Understanding Risk Appetite:

- Establish the level of risk the laboratory is willing to accept, considering its operational objectives and capabilities.

Applying a Prioritization Matrix:

- Utilize a matrix to categorize risks according to their severity and probability. This approach helps in focusing on risks with the highest potential impact on safety, compliance, and critical system functionality.

- Risks should be categorized as per time horizon, decision level in the chain of organisational command, specialist expertise required, data needs and effort to mitigate(low, medium, high).

Incorporating Best Practices:

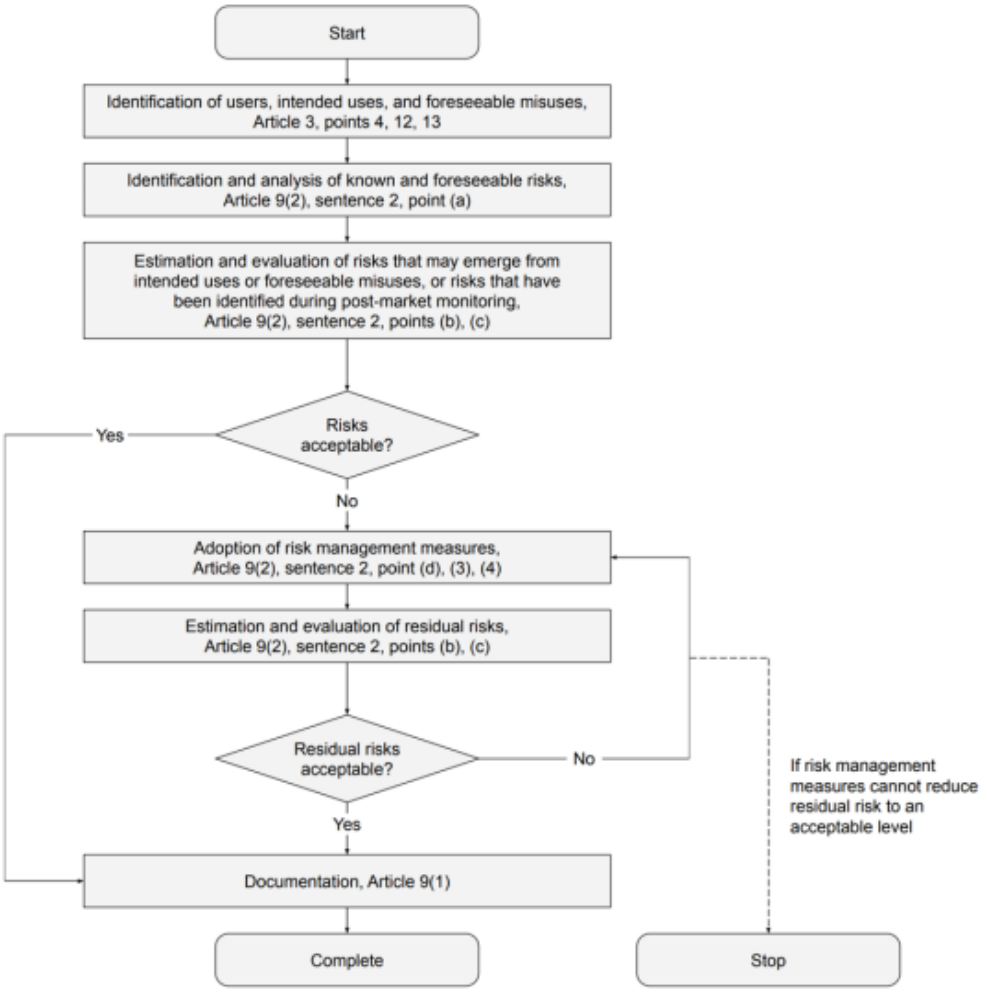
- Leverage insights from frameworks like the NIST AI RMF and ISO guidelines for risk management to ensure a comprehensive and structured approach.

Documenting and Updating Risk Priorities:

- Maintain detailed records of the risk evaluation process and outcomes. Regularly review and update the risk prioritization to reflect changes in the laboratory environment or emerging risks.

Example: Prioritize the risk of biased decision-making in an AI legal advisory tool due to its high impact on legal outcomes.

4. Risk Mitigation




*Overview of Risk Management Process as described in Article 9, Schuett, 2022*

**Objective:** To develop and implement targeted risk controls, specifically addressing unique laboratory hazards like chemical risks, data integrity issues, and other operational hazards.

**Process:**

Developing Risk Controls:

- Identify control measures for each identified risk, focusing on elimination, substitution, engineering controls, administrative controls, and personal protective equipment, in that order of preference.
- For chemical hazards, establish protocols for safe handling, storage, and disposal based on material safety data sheets and relevant regulations.

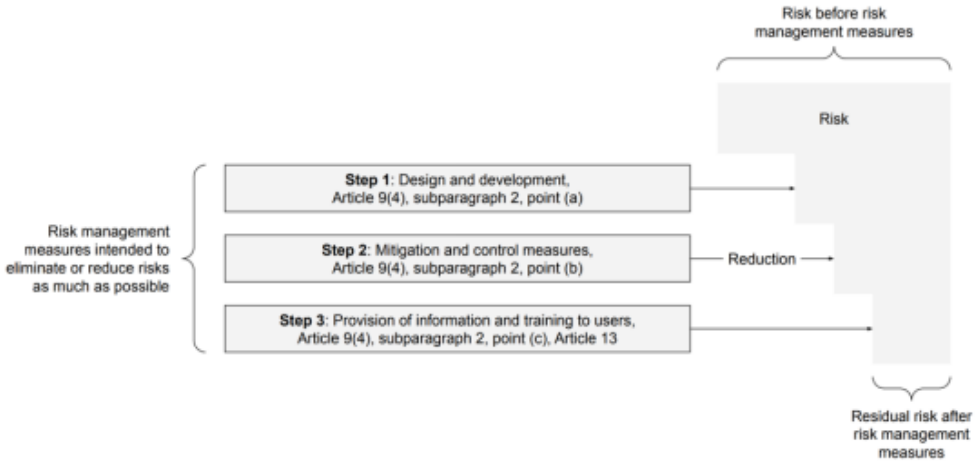
Ensuring Data Integrity:

- Implement controls for data integrity, including secure data storage, access controls, and regular data audits.
- Apply principles from ISO/IEC 27001 for information security management to protect sensitive lab data.

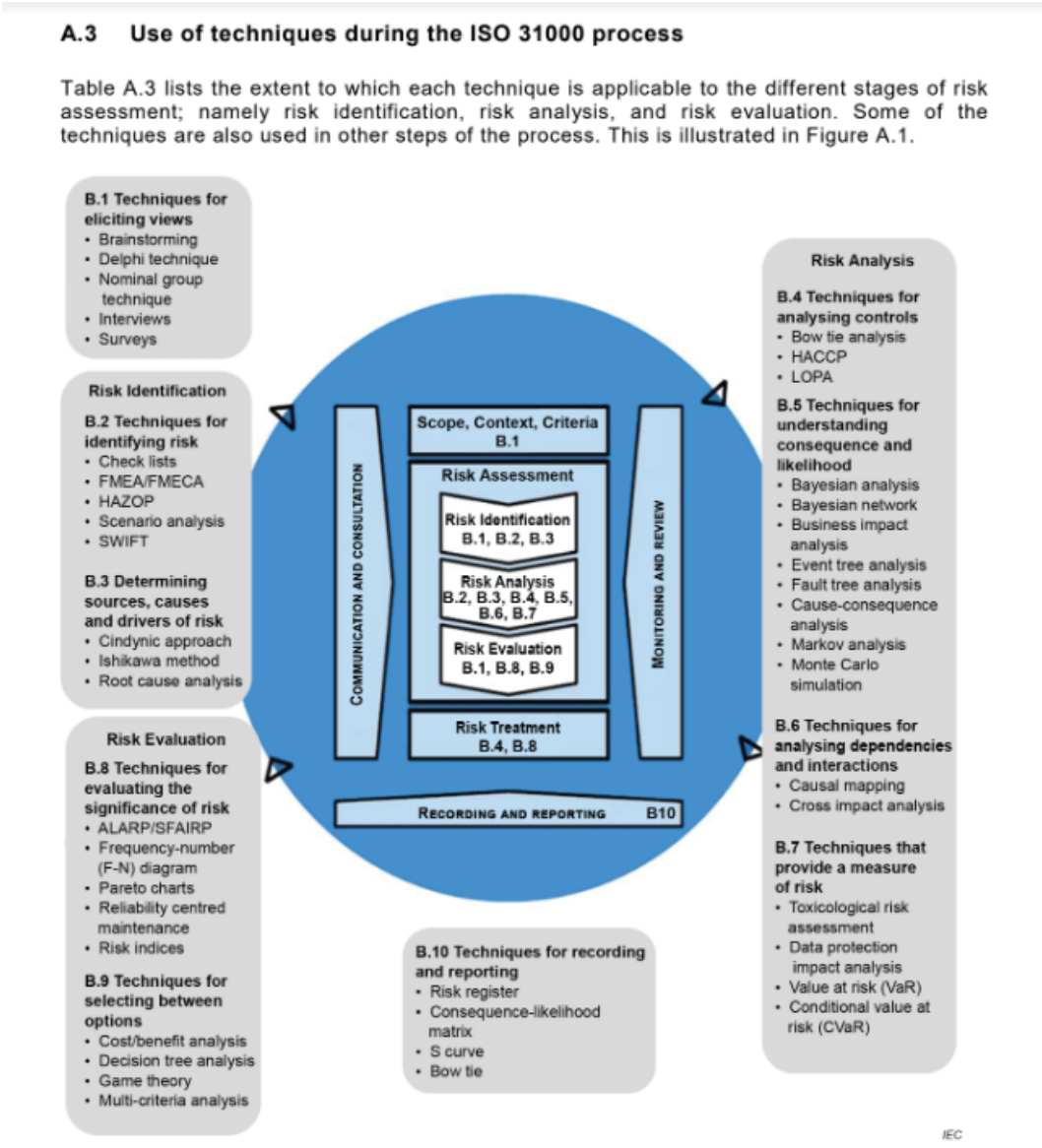
Regular Training and Drills:

- Conduct regular training sessions for lab personnel on risk mitigation measures.
- Organize drills and simulations for emergency response, particularly for handling chemical spills or data breaches.

By systematically developing and implementing these risk controls, laboratories can significantly mitigate potential hazards and ensure a safe and compliant working environment.




*Overview of Risk Management Measures in Article 9(4), Schuett, 2022*




*Different techniques for risk evaluation, identification, analysis and treatment as per ISO 31000*

5. Monitoring and Reporting:

**Objective:** Implement a robust system for continuous monitoring and reporting of compliance performance with respect to the EU AI Act, particularly aligning with the requirements in its annexures and Articles 61-63.

**Process:**

Establishing Monitoring Mechanisms:

- Set up continuous monitoring systems using both automated tools and manual checks to track compliance with specific requirements outlined in the EU AI Act, including those in the annexures.

- Incorporate ISO standards like ISO 31000 for risk monitoring and ISO 27001 for information security management.

Reporting System:

- Develop a structured process for reporting compliance issues and incidents of non-compliance.
- Align reporting mechanisms with the requirements of Articles 61-63 of the EU AI Act, ensuring timely notification to relevant authorities in case of serious incidents or breaches.

Documentation and Record-Keeping:

- Keep meticulous records of all monitoring and reporting activities, maintaining compliance documentation as per ISO standards.

6. Documentation and Record Keeping:

**Objective:** To effectively log performance monitoring of their risk management system,

**Process:**

Define Key Performance Indicators (KPIs):

- Identify specific, measurable indicators that reflect the effectiveness of the risk management system. For instance, the frequency of risk incidents, the time taken to respond to and resolve risk issues, and the number of compliance breaches.

Implement Monitoring Tools:

- Utilize software tools and systems to track these KPIs in real-time. These tools should be capable of generating alerts and reports, allowing for timely responses to potential risk management failures.

Regular Data Collection:

- Establish a routine for collecting data related to each KPI. This could involve automated data collection methods and regular manual checks to ensure accuracy.

Analysis and Reporting:

- Analyze the collected data to identify trends, areas for improvement, and the overall effectiveness of the risk management strategies. Periodically compile this data into comprehensive reports for internal review and external audits if necessary. Article 20 necessitates the maintenance of logs of high-risk AI systems for 6 months.

Documentation:

- Maintain a logbook or digital record that chronologically documents all performance monitoring activities, findings, and actions taken. This record should be easily accessible and securely stored.

Review and Adjust KPIs:

- Regularly review the relevance and effectiveness of the KPIs themselves, adjusting them as necessary to align with evolving risk landscapes, technological advancements, and changes in regulatory requirements.

7. Compliance Policy Development:

**Objective:** To craft a detailed compliance policy that not only aligns with legal requirements, including the EU AI Act, but also integrates with the organization's strategic objectives and risk profile.

**Process:**

Drafting the Policy:

- Clearly state the objective of compliance, referencing specific clauses of the EU AI Act and other relevant legislation such as GDPR for data protection.
- Define ethical standards, particularly those relevant to AI such as transparency, fairness, and accountability.

Alignment with Organizational Goals:

- Tailor the policy to support the organization's strategic objectives, ensuring that compliance also aids in achieving business goals.
- Incorporate risk management strategies that align with the organization's risk tolerance, as per ISO 37301.

Detailed Roles and Responsibilities:

- Outline specific roles and responsibilities for compliance, ensuring clarity in who is accountable for different aspects of the AI Act's provisions.

Regular Policy Review and Update:

- Establish a schedule for regular policy reviews and updates to ensure ongoing relevance and alignment with evolving laws and standards.

8. Leadership and Organizational Roles:

**Objective:** Define clear roles and responsibilities within the organization to ensure effective compliance management.

**Process:**

- Assign specific compliance roles, detailing responsibilities and authority levels.
- Engage top management, securing their commitment and leadership in fostering a compliance culture.
- Develop a culture of compliance throughout the organization, emphasizing the importance of adherence to policies and regulations.
- Utilize ISO 37301 guidelines to reinforce the structure and effectiveness of organizational roles in compliance management.

9. Continuous Improvement:

**Objective:** To establish a culture of continuous improvement in the compliance management system, ensuring it remains effective and up-to-date.

**Process:**

Scheduled Reviews:
- Set up regular intervals for reviewing the entire compliance management system.
- Assess the system's effectiveness in meeting current regulatory requirements and organizational objectives.

Feedback Integration:
- Create channels for receiving feedback from employees, regulatory bodies, and other stakeholders.
- Analyze feedback for insights and potential areas of improvement.

Lessons Learned Application:
- Document lessons learned from compliance incidents, audits, and reviews.
- Systematically incorporate these lessons into system enhancements to prevent future occurrences.

*Note:* Article 61 necessitates that high-risk AI system providers have a post-deployment monitoring plan which covers risks.

By following this process, the compliance management system will continually evolve, becoming more efficient and robust in handling the dynamic landscape of regulatory compliance.

## 5. Conclusion:

This implementation guidance equips providers with the necessary tools to efficiently manage risks in high-risk AI systems, ensuring compliance with the EU AI Act and relevant ISO standards. Its focus on rigorous monitoring, continuous improvement, and compliance fosters a safer and more responsible AI environment. As AI technology evolves, this guidance will remain a vital resource for navigating the complexities of AI risk management and regulatory adherence.

## 6. References:

ISO 31000. ISO, www.iso.org/iso-31000-risk-management.html.
ISO 37301:2021." ISO, www.iso.org/standard/75080.html.
Risk management in the Artificial Intelligence Act, Jonas Schuett, https://arxiv.org/ftp/arxiv/papers/2212/2212.03109.pdf
ISO 23984
ISO 27005
ISO/IEC 51
ISO 31010
NIST AI Risk Management Framework

## Appendix 1:

1.‘artificial intelligence system’ (AI system) means a system that is designed to operate with elements of autonomy and that, based on machine and/or human-provided data and inputs, infers how to achieve a given set of objectives using machine learning and/or logic- and knowledge based approaches, and produces system-generated outputs such as content (generative AI systems), predictions, recommendations or decisions, influencing the environments with which the AI system interacts;

1A. ‘life cycle of an AI system’ means the duration of an AI system, from design through retirement. Without prejudice to the powers of the market surveillance authorities, such retirement may happen at any point in time during the post-market monitoring phase upon the decision of the provider and implies that the system may not be used further. An AI system lifecycle is also ended by a substantial modification to the AI system made by the provider or any other natural or legal person, in which case the substantially modified AI system shall be considered as a new AI system.

1B. ‘general purpose AI system’ means an AI system that - irrespective of how it is placed on the market or put into service, including as open-source software - is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection, question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems;

2. ‘provider’ means a natural or legal person, public authority, agency or other body that develops an AI system or that has an AI system developed and places that system on the market or puts it into service under its own name or trademark, whether for payment or free of charge;

3A. ‘small and medium-sized enterprise’ (SMEs) means an enterprise as defined in the Annex of Commission Recommendation 2003/361/EC concerning the definition of micro, small and medium-sized enterprises;

4. ‘user’ means any natural or legal person, including a public authority, agency or other body, under whose authority the system is used;

5. ‘authorised representative’ means any natural or legal person physically present or established in the Union who has received and accepted a written mandate from a provider of an AI system to, respectively, perform and carry out on its behalf the obligations and procedures established by this Regulation;

5A. ‘product manufacturer’ means a manufacturer within the meaning of any of the Union harmonisation legislation listed in Annex II;

6. ‘importer’ means any natural or legal person physically present or established in the Union that places on the market an AI system that bears the name or trademark of a natural or legal person established outside the Union;

7. ‘distributor’ means any natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market;

8. ‘operator’ means the provider, the product manufacturer, the user, the authorised representative, the importer or the distributor;

9. ‘placing on the market’ means the first making available of an AI system on the Union market;

10. ‘making available on the market’ means any supply of an AI system for distribution or use on the Union market in the course of a commercial activity, whether in return for payment or free of charge;

11. ‘putting into service’ means the supply of an AI system for first use directly to the user or for own use in the Union for its intended purpose;

12. ‘intended purpose’ means the use for which an AI system is intended by the provider, including the specific context and conditions of use, as specified in the information supplied by the provider in the instructions for use, promotional or sales materials and statements, as well as in the technical documentation;

13. ‘reasonably foreseeable misuse’ means the use of an AI system in a way that is not in accordance with its intended purpose, but which may result from reasonably foreseeable human behaviour or interaction with other systems;

14. ‘safety component of a product or system’ means a component of a product or of a system which fulfils a safety function for that product or system or the failure or malfunctioning of which endangers the health and safety of persons or property;

15. ‘instructions for use’ means the information provided by the provider to inform the user of in particular an AI system’s intended purpose and proper use;

16. ‘recall of an AI system’ means any measure aimed at achieving the return to the provider or taking it out of service or disabling the use of an AI system made available to users;

17. ‘withdrawal of an AI system’ means any measure aimed at preventing an AI system in the supply chain being made available on the market;

18. ‘performance of an AI system’ means the ability of an AI system to achieve its intended purpose;

19. ‘conformity assessment’ means the process of verifying whether the requirements set out in Title III, Chapter 2 of this Regulation relating to a high-risk AI system have been fulfilled;

20. ‘notifying authority’ means the national authority responsible for setting up and carrying out the necessary procedures for the assessment, designation and notification of conformity assessment bodies and for their monitoring;

21. ‘conformity assessment body’ means a body that performs third-party conformity assessment activities, including testing, certification and inspection;

22. ‘notified body’ means a conformity assessment body designated in accordance with this Regulation and other relevant Union harmonisation legislation;

23. ‘substantial modification’ means a change to the AI system following its placing on the market or putting into service which affects the compliance of the AI system with the requirements set out in Title III, Chapter 2 of this Regulation, or a modification to the intended purpose for which the AI system has been assessed. For high-risk AI systems that continue to learn after being placed on the market or put into service, changes to the high-risk AI system and its performance that have been pre-determined by the provider at the moment of the initial conformity assessment and are part of the information contained in the technical documentation referred to in point 2(f) of Annex IV, shall not constitute a substantial modification.

24. ‘CE marking of conformity’ (CE marking) means a marking by which a provider indicates that an AI system is in conformity with the requirements set out in Title III, Chapter 2 or in Article 4b of this Regulation and other applicable Union legal act harmonising the conditions for the marketing of products (‘Union harmonisation legislation’) providing for its affixing;

25. ‘post-market monitoring system’ means all activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems they place on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions;

26. ‘market surveillance authority’ means the national authority carrying out the activities and taking the measures pursuant to Regulation (EU) 2019/1020;

27. ‘harmonised standard’ means a European standard as defined in Article 2(1)(c) of Regulation (EU) No 1025/2012;

28. ‘common specification’ means a set of technical specifications, as defined in point 4 of Article 2 of Regulation (EU) No 1025/2012 providing means to comply with certain requirements established under this Regulation;

29. ‘training data’ means data used for training an AI system through fitting its learnable parameters;

30. ‘validation data’ means data used for providing an evaluation of the trained AI system and for tuning its non-learnable parameters and its learning process, among other things, in order to prevent overfitting; whereas the validation dataset can be a separate dataset or part of the training dataset, either as a fixed or variable split;

31. ‘testing data’ means data used for providing an independent evaluation of the trained and validated AI system in order to confirm the expected performance of that system before its placing on the market or putting into service;

32. ‘input data’ means data provided to or directly acquired by an AI system on the basis of which the system produces an output;

33. ‘biometric data’ means personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, such as facial images or dactyloscopic data;

34. ‘emotion recognition system’ means an AI system for the purpose of identifying or inferring psychological states, emotions or intentions of natural persons on the basis of their biometric data;

35. ‘biometric categorisation system’ means an AI system for the purpose of assigning natural persons to specific categories on the basis of their biometric data;

36. ‘remote biometric identification system’ means an AI system for the purpose of identifying natural persons typically at a distance, without their active involvement, through the comparison of a person’s biometric data with the biometric data contained in a reference data repository;

37. ‘real-time’ remote biometric identification system’ means a remote biometric identification system whereby the capturing of biometric data, the comparison and the identification all occur instantaneously or near instantaneously;

38. ‘publicly accessible space’ means any publicly or privately owned physical place accessible to an undetermined number of natural persons regardless of whether certain conditions or circumstances for access have been predetermined, and regardless of the potential capacity restrictions;

39. ‘law enforcement authority’ means:

any public authority competent for the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security; or
any other body or entity entrusted by Member State law to exercise public authority and public powers for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security;

40. ‘law enforcement’ means activities carried out by law enforcement authorities or on their behalf for the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security;

41. ‘AI regulatory sandbox’ means a concrete framework set up by a national competent authority which offers providers or prospective providers of AI systems the possibility to develop, train, validate and test, where appropriate in real world conditions, an innovative AI system, pursuant to a specific plan for a limited time under regulatory supervision.

42. ‘national competent authority’ means any of the following: the notifying authority and the market surveillance authority. As regards AI systems put into service or used by EU institutions, agencies, offices and bodies, the European Data Protection Supervisor shall fulfil the responsibilities that in the Member States are entrusted to the national competent authority and, as relevant, any reference to national competent authorities or market surveillance authorities in this Regulation shall be understood as referring to the European Data Protection Supervisor;

43. ‘informed consent’ means a subject's free and voluntary expression of his or her willingness to participate in a particular testing in real world conditions, after having been informed of all aspects of the testing that are relevant to the subject's decision to participate; in the case of minors and of incapacitated subjects, the informed consent shall be given by their legally designated representative;

44. ‘serious incident’ means any incident or malfunctioning of an AI system that directly or indirectly leads to any of the following:

the death of a person or serious damage to a person’s health;
a serious and irreversible disruption of the management and operation of critical infrastructure;
breach of obligations under Union law intended to protect fundamental rights;
serious damage to property or the environment.

45. ‘critical infrastructure’ means an asset, system or part thereof which is necessary for the delivery of a service that is essential for the maintenance of vital societal functions or economic activities within the meaning of Article 2(4) and (5) of Directive …../….. on the resilience of critical entities;

46. ‘personal data’ means data as defined in point (1) of Article 4 of Regulation (EU) 2016/679;

47. ‘non-personal data’ means data other than personal data as defined in point (1) of Article 4 of Regulation (EU) 2016/679;

48. ‘testing in real world conditions’ means the temporary testing of an AI system for its intended purpose in real world conditions outside of a laboratory or otherwise simulated environment with a view to gathering reliable and robust data and to assessing and verifying the conformity of the AI system with the requirements of this Regulation; testing in real world conditions shall not be considered as placing the AI system on the market or putting it into service within the meaning of this Regulation, provided that all conditions under Article 53 or Article 54a are fulfilled;

49. ‘real world testing plan’ means a document that describes the objectives, methodology, geographical, population and temporal scope, monitoring, organisation and conduct of testing in real world conditions;

50. ‘subject’ for the purpose of real-world testing means a natural person who participates in testing in real world conditions.

## 2. Discussion and Conclusion

Let’s add a few citations: (Alon, 2006; Goh et al., 2021; Lindner et al., 2023; Olah et al., 2020; Weiss et al., 2021). Here, we use the Zotero software with the Zotero add-on for Chrome and Google Docs for citation management but you can also write them manually (for significantly more hassle).

## 3. References


Alon, U. (2006). *An Introduction to Systems Biology: Design Principles of Biological Circuits* (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420011432

Goh, G., Cammarata, N., Voss, C., Carter, S., Petrov, M., Schubert, L., Radford, A., & Olah, C. (2021). Multimodal Neurons in Artificial Neural Networks. *Distill*, *6*(3), 10.23915/distill.00030. https://doi.org/10.23915/distill.00030

Lindner, D., Kramár, J., Rahtz, M., McGrath, T., & Mikulik, V. (2023). *Tracr: Compiled Transformers as a Laboratory for Interpretability*

(arXiv:2301.05062). arXiv. http://arxiv.org/abs/2301.05062
Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom In: An Introduction to Circuits. *Distill*, *5*(3), 10.23915/distill.00024.001. https://doi.org/10.23915/distill.00024.001
Weiss, G., Goldberg, Y., & Yahav, E. (2021). *Thinking Like Transformers* (arXiv:2106.06981). arXiv. http://arxiv.org/abs/2106.06981

## 4. Appendix

Add any extra content you wish here! Unrestricted.