GPA

Global Privacy Assembly

GPA AIWG

RISKS FOR RIGHTS AND FREEDOMS OF INDIVIDUALS POSED BY ARTIFICIAL INTELLIGENCE SYSTEMS - PROPOSAL FOR A GENERAL RISK MANAGEMENT FRAMEWORK

AIWG Action Point n.6

SUMMARY

INTRODUCTION

As stated in the Declaration on Ethics and Data Protection in Artificial Intelligence[1], the use of artificial intelligence (AI) engages significant privacy, data protection and ethical concerns affecting both individuals and society at large that may involve the curtailing fundamental rights and freedoms. The objective of this document is to draw a general framework for identifying the privacy risks posed by AI systems. It provides a general overview of the primary risks of AI systems to individual's privacy based on internationally recognized privacy principles,[2] and outlines an overall strategy to help prevent harm to individuals and society.

This framework is to provide an outline of risk factors in AI development and deployment. Organizations remain responsible for the applicable Data Protection and Privacy requirements.

[1] Declaration on Ethics and Data Protection in Artifical Intelligence - 40th International Conference of Data Protection and Privacy Commissioners - Tuesday 23rd October 2018, Brussels (see: http://globalprivacyassembly.org/wp-content/uploads/2018/10/20180922_ICDPPC-40th_AI-Declaration_ADOPTED.pdf).

[2] This includes the 2019 OECD Recommendation on AI that provides a set of internationally-agreed principles and recommendations that can promote an AI-powered crisis response that is trustworthy and respects human-centered and democratic values (see: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449).

AI is a key technology that is fundamentally transforming our economies and societies. It has great potential in helping us address some of the biggest challenges we face in our world today, such as healthcare delivery, supply chain management, and energy consumption. However, when AI is used as an advanced form of processing personal information, it is important to clarify how traditional privacy principles and practices should be applied.

AI systems can pose risks to the rights and freedoms of natural persons in various ways, some of which are difficult to identify, foresee and substantiate. The benefits of efficiency and scale are often traded for decisions that are difficult to understand or interrogate, including when such systems are used to make recommendations and predictions, draw inferences, compose profiles, and influence individuals. This creates the need for special considerations to ensure AI systems respect individuals' privacy and other fundamental rights, including self-determination, anti-discrimination and human dignity.

Consideration of the principles of data protection and the values outlined in this document as part of an organization's risk-management process will be a key factor in the development of trusted AI technologies.

A RISK MANAGEMENT FRAMEWORK FOR AI, ETHICS AND DATA PROTECTION

The Declaration calls for continued attention, vigilance, and accountability of the whole chain of actors and stakeholders, including supervisory authorities**,** individuals and other parties as appropriate, on the potential outcomes and impacts of AI systems.

Each of these relevant actors can make a difference in the way AI systems are designed, developed and deployed. They should carefully consider the complexity of the AI systems (including their training and input data, hardware and software parameters, etc.) their application domain, and the variety of the specific contexts and environments in which systems are applied and operated.

Data protection is largely driven by a risk-based approach. In order to protect privacy and other fundamental rights and freedoms, as well as the safety and security of individuals, a risk management approach and related best practices are key instruments for actors in the AI value chain to be accountable, ensuring continued attention and vigilance. As also pointed out in the Declaration, *"those risks and challenges may affect individuals and society, and that the extent and nature of potential consequences are currently uncertain".* The risk management process can anticipate possible harm to individuals and it's to be intended as an iterative process to be activated along all the steps of a system's lifecycle, when it is designed, developed, deployed, used, as well as when relevant changes in the system or the application context occur.

The figure below presents a general model for risk management in AI systems which reflects the reality that several actors/stakeholders with different tasks and responsibilities are involved in the development and deployment of AI systems.

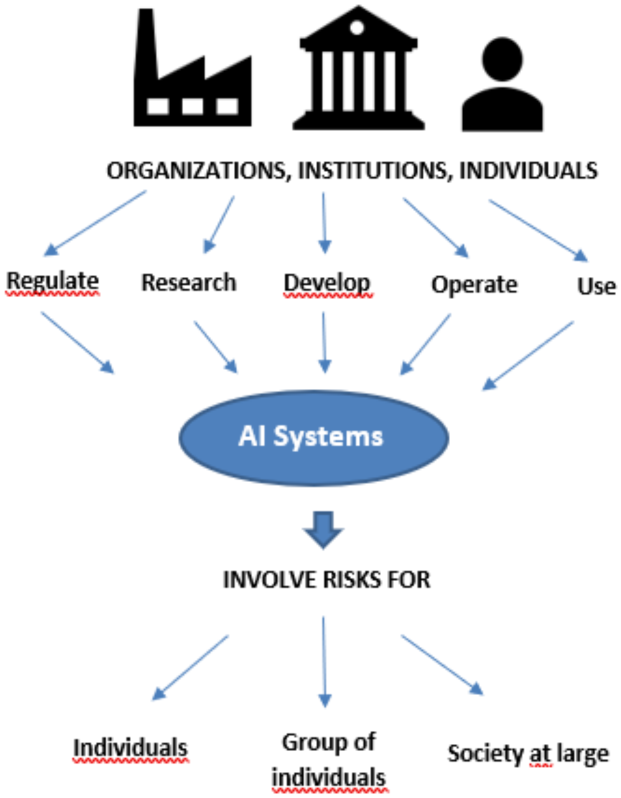


Fig. 1 AI Risk Management general model

MANAGING THE RISK OF AI SYSTEMS

The development and use of AI must be preceded by a comprehensive data protection impact assessment. According to Risk Management standards[3] "*The purpose of risk management is the creation and protection of value - It improves performance, encourages innovation and supports the achievement of objectives*".

The first necessary step of putting in place a sound risk management process is to identify the *value* that needs to be protected. Taking a human-centric approach, the design, development and application of AI systems should focus on people's welfare. As stated in the Declaration: "*any creation, development and use of Artificial Intelligence systems shall fully respect human rights, particularly the rights to the protection of personal data and to privacy, as well as human dignity, non-discrimination and fundamental values, and shall*

[3] ISO 31000:2018 - Risk management Guidelines.

provide solutions to allow individuals to maintain control and understanding of Artificial Intelligence systems.”

The following aspects in particular must be considered for this value-based approach:

1) Fairness and lawfulness:

- The processing of personal data must be lawful. Organizations must identify their lawful authority to collect, use, or disclose the personal information that they intend to process as part of a proposed AI system. This can include verifying any data relied upon for training AI systems were collected in accordance with data protection law.
- Compliance with the general data protection principles is an indispensable prerequisite for the use of AI that is permissible under data protection law.
- The processing of personal data must be fair and lead to fair outcomes. This includes collecting data in a way that does not deceive or mislead individuals, consideration of individuals’ reasonable expectations, and any adverse impacts and foreseeable risk of harm to individuals, groups of individuals, or society at large.
- The purposes for which AI is used must be in good faith and not seek or result in the exploitation of human vulnerabilities, exacerbation of inequality, discrimination, increase societal divisions, or result in unjustified material or immaterial harm.
- Interference with the right to informational self-determination requires a legal basis that sufficiently limits the use of data to specific purposes. Specific legal authority is required in particular if AI is to be used for law enforcement and security purposes.

2) Transparency and explainability:

- Individuals should be informed of when AI will be used to process their personal information for an outcome, including to reach decisions, make recommendations and predictions, draw inferences, compose profiles, or influence individuals.
- Guaranteeing the right to informational self-determination, notably by ensuring that individuals are always informed appropriately when they are interacting directly with an AI system or when they provide personal data to be processed by such systems. Inferential and predictive data concerning an individual is personal information, and hence subject to data protection rules; including transparency. It must be possible to explain the specific functions of the AI system to affected individuals using plain language, and in a manner that is clear and understandable. Although this may be particularly challenging both for developers and organizations when outsourcing or adopting AI systems, explainability should be considered as a core characteristic of AI systems.

- Where an AI system is used to take automated decisions about individuals, they should be provided with a meaningful explanation of how the decision was reached which at a minimum allows them to understand:
  - (i) the nature of the decision they are subject to and the relevant personal information it relied upon, and
  - (ii) the rules that define the processing and the decisions key parameters

  to be able to exercise their rights appropriately.
- Trade secrets should not be invoked to deny sufficient information and explanations. Organizations should at least provide the individual with:
  - the type of information collected or used in creating the profile or making the automated decision;
  - why the information is relevant; and
  - what the likely impact is going to be.
- Where AI is used to draw or create any new personal information, such as inferences or make predictions, profiling, categorization, or to provide recommendations to an individual, data subjects must be able to exercise their right to access such information.
- It must be possible to explain how the AI-system was trained. In principle, it must also be clear which data has been processed and which models and systems have been used.
- In some narrow cases, such as the detection of fraud or law enforcement, transparency and explainability obligations toward the individual concerned could be restricted as far as necessary in order not to jeopardize the legitimate purpose(s) at stake. However, these obligations should not prejudice but enable compliance with other legal requirements such as those in the context of financial services regulation or law enforcement.
- AI systems, must be able to be comprehensively overseen by data protection supervisory authorities and relevant regulators, including the ability to inspect the underlying data used by AI, and the models or systems themselves.

3) Recourse and redress:

- Where people are affected by the consequences of a decision based on an AI-system, they must receive sufficient information to be able to exercise their rights appropriately.
- Individuals must have the opportunity through easily accessible information and complaint channels to question, challenge or contest automated decisions produced by AI systems based on personal information, if necessary. A recourse mechanism should be established to allow for human intervention and review.

- In some jurisdictions people have the right not to be subject, in principle, to a decision based solely on automated processing which significantly affects them (for example in Art. 22 GDPR), and have an ability to seek human intervention, express their point of view, and contest the decision.

4) Data minimization and storage limitation:

- Whenever possible, the least amount of personal information should be used to train and operate AI systems, including the use of anonymized or synthetic data. Any storage of personal data should not be kept longer than is necessary.

5) Purpose limitation of data processing:

- Only data that serves the directly specified purpose may be used, unless an exception applies. Data that has already been collected for other purposes may be processed for further purposes in some jurisdictions if these secondary purposes are compatible with the original purpose of the data collection. Some jurisdictions also have research exceptions that can allow for the training of some types of AI systems. Otherwise, a standalone authority must be relied upon or consent obtained must be used as applicable.

6) Accuracy of data and data quality:

- Personal data are required to be correct and up to date (current), avoiding any inaccuracy in collection and processing. Every reasonable step must be taken to ensure that personal data that relates to error-free records can be used as a reliable source of information. In data management, data accuracy is the first and critical component/standard of the data quality framework.
- AI-based or assisted decisions must be regularly checked for non-discrimination. In doing so, routine reviews for bias in algorithmic decision-making systems that may threaten fundamental rights should be conducted. Data can be, along with model decisions or context of deployment, one of the sources of bias. Therefore, incorrect or unrepresentative data should not be used to train models and should be deleted or corrected if necessary.

7) Accountability and liability:

- Organizations remain responsible for adverse impacts caused by their AI systems, and should appoint qualified individuals to assess AI conformity with data protection and privacy laws and frameworks.
- Accountability of all relevant stakeholders to individuals should be identified, such as a liability system to compensate for any damage caused by AI-systems.

8) Ensuring data security - availability, integrity and confidentiality:

- It’s essential to ensure data security. The protection of training, input and output data involved in the development and operation of AI systems is necessary to ensure the integrity and correct functionality of AI systems and hence protection to individuals. The technical and organizational security measures should always be appropriate to the current state of the art.

9) Consideration of ethical aspects:

- The starting point for ethical considerations is the question of how we want to live in a world with AI-driven or supported decision-making. An important consideration is that AI systems are useful tools of huge potential, but they are not moral agents who can act responsibly. Only humans can be responsible. It is therefore important to democratically interrogate the changes that arise from developments in the field of AI.
- AI requires a detailed empirical review and a comprehensive socio-political discussion to clarify, on the one hand, the impact of this technology on people’s freedoms and, on the other hand, to determine the necessity of its use especially for law enforcement and security purposes.
- If AI applications are to benefit humans they need to abide by the legal frameworks set out to protect them and their fundamental rights. The core area of private life and the guarantee of human dignity must not be undermined by the use of AI. In particular, the use of AI for "social scoring" can lead to discrimination and contradict these fundamental values. Depending on the application context, procedures that use AI to automatically recognise human characteristics (e.g. faces, gait or other biometric characteristics) can also be highly problematic, for example when used in publicly accessible spaces.
- An ethical lens should be applied to the practice of drawing group-level statistical inferences using big data analysis. The invasiveness of group-level inferences is an important ethical consideration in AI systems. Just because you may have the ability to draw correlations about an individual by relying on analysis of the behaviour of specific ad-hoc groups of individuals in big datasets, does not mean you should always apply these group-level inferences. This type of analysis also runs the risk of reinforcing stereotypes, bias, discrimination, and historical injustices. It could deprive an individual from paving their own pathway and breaking free from societal preconceptions. Given that AI has the potential to infer unexpressed characteristics about an individual using such group-based analysis, and to use such information to influence their thought and behaviour, ethical

limitations should be drawn where personal information is used to undermine autonomy or rational capacity for decision-making. This includes exploiting individuals' mental health, addiction, and other susceptibilities.

- It is also important to acknowledge that the ethical values of the developers of AI systems may influence the systems they create: formulating ethical principles in connection with AI and communicating them transparently can contribute to the ethical adoption of human-centric AI.

Among others, the following ethical principles are considered particularly relevant:

- Non-maleficence (security, physical and metal safety).
- Beneficence (wellbeing, preserve human dignity, equality, sustainability and solidarity).
- Justice and fairness of outcome, including non-discrimination.
- Liability
- Whistleblowing protections and processes
- Autonomy (self-determination, freedom of choice).

THE RISK MANAGEMENT PROCESS FOR AI SYSTEMS

Actors and Stakeholders

Each of the following actors and stakeholders have a role to play and should exercise their power and influence to promote accountability for fair and responsible development and use of AI systems, respectful of the rights and freedoms of individuals and the society at large also through the design, planning and implementation of the technical and organizational measures as above identified.

- Regulators (Legislators, Public Authority with governance/enforcement powers)
- Researchers (Academia, Public and private research entities)
- Standards Organizations (developing standards and best practices)
- Producers and Providers (Designers and developers of algorithms, software and related data structure for machine learning human interfaces and actuators, system integrator, producers of system and providers of services, data providers - (Combine Designers, Producers, Operators,
- End users (as users of AI systems)

Even though the overall responsibility for the use of the AI system always lies with the controller who is responsible for this processing in terms of data protection law, it is essential that all actors make their contribution to the responsible use of AI.

The risk management process as described in the following sections requires that each relevant actor, considering the specific activity performed, should evaluate the "context" pertinent to the specific AI system and identify, analyse, evaluate the risks and implement appropriate measure to mitigate risks where appropriate. It is important to note that what is required is a continuous iterative process throughout the lifecycle of an AI system, as many systems or the environment they are deployed in are dynamic.

Context definition

AI-specific risks depend to a large extent on the deployment scenario and the parameters of the AI system. For each AI system (or "c*lass*" of AI system where is possible to group AI systems with a certain degree of commonality, e.g. healthcare-medical diagnosis, image or speech recognition, etc.) both the characteristics (the scope and related functionalities, the assumptions made in problem formulation and proxies used, the origin, type and nature of personal data processed including the relevant processing operations applied) and context of application must be considered

In order to identify the risks to rights and freedoms of AI-relevant procedures, it is necessary to distinguish what kind of support the AI systems are intended to provide and the context of their use. For example, the risks of AI systems that are only intended to provide one-off retail recommendations on individual products must be assessed differently than the risks of AI-system decisions in more impactful areas, such as health diagnoses.

Defining the context is fundamental in order to identify the relevant stakeholders and the potential risks for individuals. The social, cultural, ethical values characterizing the application contexts and the individuals affected by AI systems should be carefully considered to identify the possible origin of adverse consequences (risks).

Risk identification

Depending on the specific characteristics of an AI system and the *context* in which it is deployed, natural persons may be subject to different risks. The origins of such risks are connected to the characteristics, approaches and techniques used to develop of AI system. Certain autonomous systems and Machine Learning (ML) applications may carry more easily identifiable risks that should be evaluated in relation to the specific application context and the categories of individuals affected by their outputs.

For example, the well-known problems of *bias* in algorithmic systems as well as the quality of the training data should be carefully considered. In this case, it's worth paying particular attention to the violation of the data minimization principle applied to training data. As mentioned earlier, to avoid discrimination, it is important to carefully select the training data used and to continuously monitor the AI system's output for discriminatory or erroneous results, since the maximization of data collected in order to reduce bias is not in and of itself a 'silver-bullet' solution. Acknowledging that AI systems can also lead to discriminatory effects that are not immediately evident or may emerge as begin to interact with the environment, adopting an adequate, ongoing risk management process is imperative. Moreover, risks to individuals are to be considered also in relation to possible harms originated by AI outputs or assisted decisions affecting them as members of a group or the society at large. This is the case of negative impacts on the environment or other consequences impacting society and influencing individual wellbeing and fundamental rights.

The main categories of risks that should be carefully considered are:

- Violation of fundamental ethical principles, including beneficence and non-maleficence of AI, autonomy and self-determination of human beings.
- Lack of transparency on the objectives, expected outputs, the inner logic of algorithms and of any component of AI system affecting rights and freedoms of individuals.
- Violation of fundamental principles of personal data protection: fair and legitimate processing and outcomes, purpose specification, transparency, proportionality and necessity, accuracy, storage limitation, and security.
- Unfair discrimination of individuals due to gender, ethnic origin, political opinions, religion or philosophical beliefs, health status, sex life, personal preferences and interests, behavior, socioeconomic conditions or any other personal condition or belief. Attention need to be paid to emerging forms of discrimination.
- Deprivation of rights and freedoms of individuals, particularly of the most vulnerable persons.
- Exercising unlawful or unfair control over personal data.
- Exploitation of human fragilities and conditions (for example, mental health, addiction, etc.)
- Any potential system's condition or outcome causing material or non-material harm to individuals, including unfairly or adversely affecting individual's entitlements to services or opportunities.
- Emergence of social harm because of the use of AI based tools and techniques to:
  - unlawfully detect, monitor or profile people;
  - spread disinformation, misleading synthetic content (e.g. 'deepfakes') and other manipulated material in order to shape opinions and behaviors of individuals or groups of people.
- Negative impacts on sustainability, related to socio-economic inequalities, loss of jobs and the carbon footprint impact due to computational power required to train AI algorithms.

Determination of risk level

The assessment of the risks of an AI system should be based on the likelihood of occurrence of the event causing harm and the severity of consequences to the rights and freedoms of natural persons. The risk level should be calculated appropriately by all relevant stakeholders in relation to the role they play in the development of AI systems. In practice this means the primary responsibility for assessing and mitigating risks must rest with developers and end-users of AI systems. For identified risks, suitable measures should be examined with risks avoided as best as possible, and to guarantee the systems are in compliance with broader national and international legal frameworks.

Mitigating measures

Below are a set of measures to be considered by the accountable stakeholders to mitigate the ethical, privacy and data protection risks:

- *Risk Management*
  - Implement a sound risk management process that specifically refers to ethics, data protection issues and respect of fundamental rights of an AI system during its entire lifecycle.

- *Ethical design, implementation and operation of AI systems*
  - Identify specific requirements for ethical and fair use of the systems and for respecting fundamental rights as part of the development and operation of any AI system.

- *Security*
  - Implement technical and organizational measures and procedures – proportional to the type of system developed, the risk level of that system, the nature of personal data processed and the categories of individuals affected – to ensure that ethical principles, data protection, privacy and other fundamental rights such as freedom of expression, freedom of assembly, and information and non-discrimination, are respected both when designing the system and, on a continuous basis when the system is used and deployed.

- *Responsible development*
  - AI systems should be developed in a way that:
    - Facilitates human development and flourishing.
    - Does not obstruct or endanger individuals/groups.
    - Sets out controls on potential high-risk use cases.

        - Provides assessment and documentation related to the expected impact on individuals and society not only at the design stage but also when major upgrades are provided and during their entire life cycle as they are being retrained.
        - Adopts the privacy, ethics and security 'by design' approach as a fundamental practice to move risk mitigation upstream and not after the system is deployed, prevent harms and for an efficient and effective approach to human-centered AI development.

- *Continuous monitoring*
    - The operation of AI systems is constantly monitored by all accountable actors to ensure it remains consistent with the intended and stated scope, and to verify that personal data are used in a way that is compatible with the original purpose of their collection.

- *Information to data subjects*
    - Providing adequate information on the purpose and effects of AI systems in order to verify continuous alignment with the expectations of individuals and to enable overall human control on such systems.
    - Guaranteeing the right to informational self-determination, by ensuring that individuals are always informed appropriately:
        - When they are interacting directly with an AI system.
        - When they provide or have their personal data processed by such systems.

- *Algorithmic transparency*
    - Ensuring algorithmic transparency and the auditability of systems, while ensuring meaningfulness of the information provided. This will often mean different levels and forms of transparency and information for each relevant audience.

- *Quality vs. Maximization of training data*.
    - Ensuring the accuracy of training data sets and the application of the data minimization principle, including by using anonymized or synthetic data.

- *Foster private and public research on human-centric AI*
    - Investing in research, among others, on the following:

    - Technical and social science approaches to identify, address and mitigate biases.
    - Develop solutions for explainable AI.
    - Develop agile solutions to bring, where necessary, the AI system under human control at any stage of its operation.

- *Raise awareness of AI risks*
  - Producing specific guidance and principles in addressing biases and discrimination, and promoting individuals' and stakeholders' awareness.
  - Investing in awareness raising, education, and training in order to ensure a good level of information on and understanding of AI and its potential effects on society.

- *Develop and enforce sectoral standards and best practices*
  - Fostering collective and joint responsibility, involving the whole chain of actors and stakeholders, including through the development of sectoral standards and the sharing of best practices.

- *Governance*
  - Establishing demonstrable governance processes for all relevant actors, such as relying on trusted third parties or the setting up of independent ethics committees.
  - Supporting data protection and privacy authorities and placing them at the center of AI governance. Effective data protection supervision requires that the competent data protection supervisory authority can control AI-based data processing. For data protection supervision, all existing mechanisms of the AI system must be efficiently traceable.

- *Accountability.*
  - Promoting accountability of all relevant stakeholders including through the realization, for AI systems, of audits, continuous monitoring, impact assessments and periodic review of existing oversight mechanisms.

- *Empowering individuals to exercise their applicable data protection and privacy rights, including:*
  - the right to information,

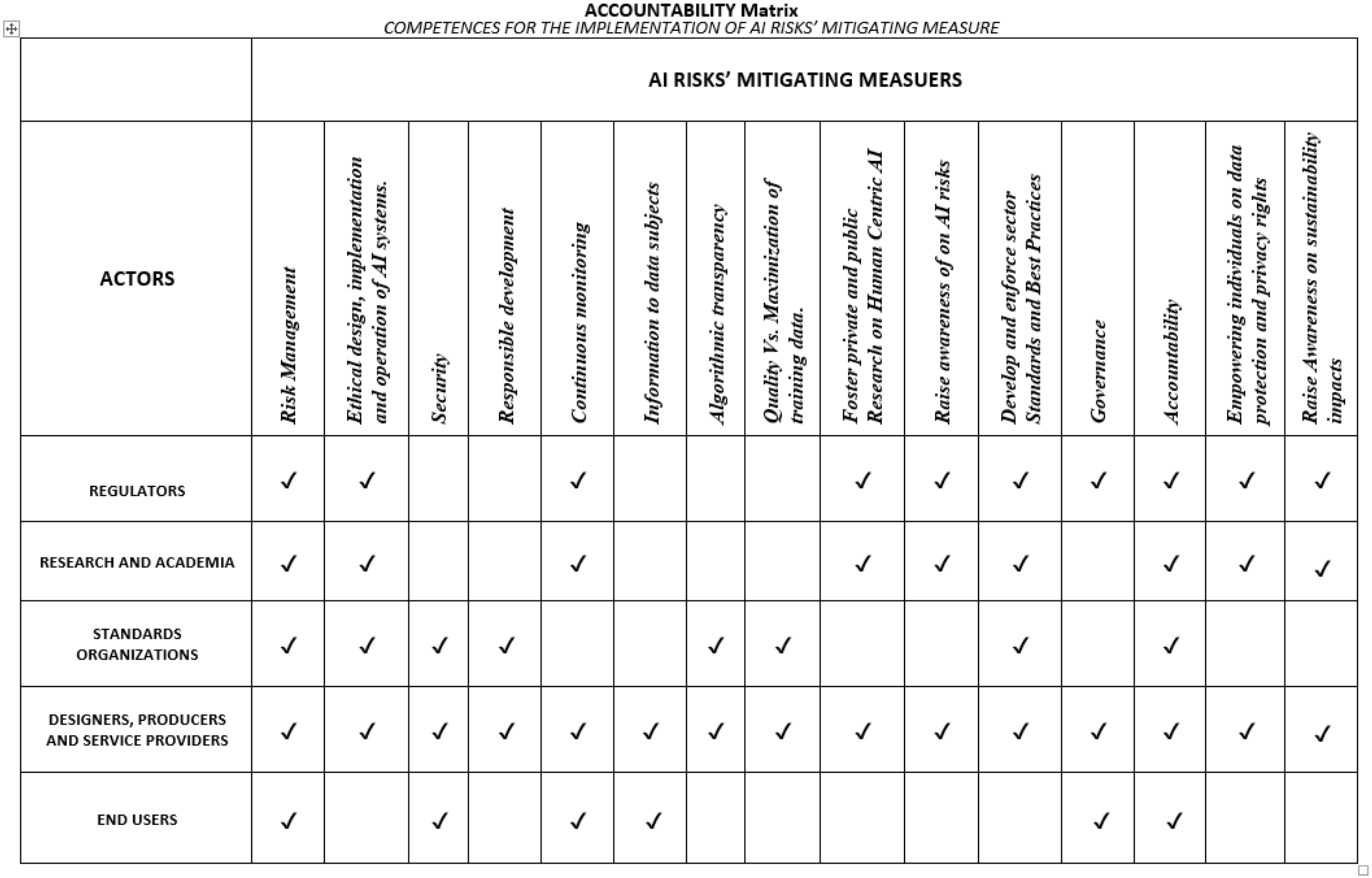
- the right to access,
- the right to object to processing,
- the right to restrict the processing,
- the right to erasure,
- the right not to be subject to a decision based solely on automated processing if it significantly affects the data subject, and
- and the right to challenge such decision.

- *Raise Awareness on sustainability impacts of AI systems*

  - Develop models to measure the impacts of AI systems on the environment, society and economy.
  - Adopt guidelines and enforce risk management practices to address the impact of AI systems on sustainability (e.g. carbon footprint of AI computational processes, consequences on employment and the exacerbation of social and economic inequalities).

Accountability Matrix

The following table shows a suggested attribution of roles and related responsibilities on the competence referred to the identification, design, planning and implementation of the above risk mitigation measures.



ACCOUNTABILITY Matrix

COMPETENCES FOR THE IMPLEMENTATION OF AI RISKS’ MITIGATING MEASURE

	AI RISKS’ MITIGATING MEASUERS														
ACTORS	*Risk Management*	*Ethical design, implementation and operation of AI systems.*	*Security*	*Responsible development*	*Continuous monitoring*	*Information to data subjects*	*Algorithmic transparency*	*Quality Vs. Maximization of training data.*	*Foster private and public Research on Human Centric AI*	*Raise awareness of on AI risks*	*Develop and enforce sector Standards and Best Practices*	*Governance*	*Accountability*	*Empowering individuals on data protection and privacy rights*	*Raise Awareness on sustainability impacts*
REGULATORS	✓	✓			✓				✓	✓	✓	✓	✓	✓	✓
RESEARCH AND ACADEMIA	✓	✓			✓				✓	✓	✓		✓	✓	✓
STANDARDS ORGANIZATIONS	✓	✓	✓	✓			✓	✓			✓		✓		
DESIGNERS, PRODUCERS AND SERVICE PROVIDERS	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
END USERS	✓		✓		✓	✓						✓	✓		