AI & Partners

EU AI Act

AI System Risk Assessment

AI & Partners Guidance

Version 1.0

Last Updated: 19 November 2023

Contents

Acronyms

Acronym	Description
AI	Artificial Intelligence
EU	European Union
RBA	Risk-Based Approach

Introduction & Terminology

Purpose, Scope, and Status of this Guidance

Identifying, assessing, and understanding AI system risks is an essential part of the implementation and development of a European Union ("EU") AI Act (the "EU AI Act") compliance regime by firms, which contains provisions to mitigate risks of harm posed by AI systems to individuals' health, safety, fundamental rights, and democracy. It assists in the prioritisation and efficient allocation of resources by firms. The results of a firm-level risk assessment, whatever its scope, can also provide useful information to stakeholders in understanding its implications. Once AI system risks are properly understood, firms may apply risk management measures in a way that ensures they are commensurate with those risks – i.e,. the risk-based approach ("RBA") – which is central to the EU AI Act.

This document is intended to provide guidance on the conduct of risk assessment at the firm level, and it relates especially to key requirements set out in the EU AI Act. In particular, it outlines general principles that may serve as a useful framework in assessing AI system risks at the firm level. The general principles contained in this paper are also relevant when conducting risk assessments of a more focussed scope, such as in assessments of a particular AI system or business function or of thematic issues (for example, transparency issues related to AI systems). All of these types of assessments (comprehensive, sectoral or thematic) carried out at the firm level may also form the basis for determining whether to apply enhanced or specific measures, simplified measures, or exemptions from risk management requirements.

The guidance in this document is not intended to explain how firms should assess risks in the context of risk-based oversight and supervision, although risk-based oversight and supervision will likely be informed by a firm-level risk assessment. Also, this guidance does not provide further explanation of RBA obligations and decisions for firms. AI & Partners intends to issue separate guidance on implementing the RBA for specific sectors and professions, and that material will be reviewed and, as necessary, modified in light of amendments to the EU AI Act.

This guidance is structured as follows:

- Section 1 lays out the purpose, scope and status of this guidance, along with an outline of the core EU AI Act obligations relevant to AI System risk assessments at any level.
- Section 2 lays out general principles that should be taken into account when conducting AI system risk assessments at the firm level.
- Section 3 presents a high-level view of the three main stages involved in the AI system risk assessment process (identification, analysis, and evaluation).
- Section 4 considers the outcome and dissemination of the risk assessment product.

- Annexes to this document contain additional information relating to AI system risk assessment.

Core EU AI Act Obligations and Decisions regarding AI System Risk Assessments

It is important that the users of this guidance have an understanding of the obligations contained in the EU AI Act. This section provides a general outline of these obligations.

Article 9 of the EU AI Act: Article 9(1) of the EU AI Act lays down a requirement for, "A risk management system shall be established, implemented, documented and maintained in relation to high-risk AI systems, throughout the entire lifecycle of the AI system. The risk management system can be integrated into, or a part of, already existing risk management procedures relating to the relevant Union sectoral law insofar as it fulfils the requirements of this article." Additionally, Article 9(2) specifies 2 that, "The risk management system shall consist of a continuous iterative process run throughout the entire lifecycle of a high-risk AI system, requiring regular review and updating of the risk management process, to ensure its continuing effectiveness, and documentation of any significant decisions and actions taken subject to this Article. It shall comprise the following steps:", taking into account, amongst other things, "(a) identification, estimation and evaluation of the known and the reasonably foreseeable risks that the high-risk AI system can pose to the health or safety of natural persons, their fundamental rights including equal access and opportunities, democracy and rule of law or the environment when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse.

Interpretative Note to Article 9 of the EU AI Act: Establishing a comprehensive risk management system for high-risk AI systems is imperative throughout their lifecycle. Integration with existing sectoral laws ensures regulatory compliance. The continuous iterative process, involving regular review and updates, is essential to adapt to evolving risks. Identification, estimation, and evaluation of potential risks, considering intended use and foreseeable misuse, are critical steps. Understanding these risks is pivotal for crafting effective mitigation strategies, safeguarding health, safety, fundamental rights, democracy, rule of law, and the environment, as well as knowing what risk category to assign to an AI system.

Key Concepts and Terms Relevant to AI System Risk Assessment

In discussing AI System risk assessment, it is useful to have a common understanding of certain key concepts and terms that will be used in this guidance. Many of these come from the area of risk management, a process commonly used in the public as well as the private sectors to help in decision-making. While many risk management concepts are usefully described elsewhere, their use in this guidance has been adapted to the particular case of assessing AI System risk at the firm level. Broadly speaking, however, risk management involves developing the appropriate measures to mitigate or reduce an assessed level of risk to a lower or acceptable level.

For the purposes of assessing AI system risk at the firm level, this guidance uses the following key concepts:

- ***Risk*** can be seen as a function of three factors: *threat*, *vulnerability* and *consequence*. An AI system risk assessment is a product or process based on a methodology, agreed by those parties involved, that attempts to identify, analyse and understand AI system risks and serves as a first step in addressing them. Ideally, a risk assessment, involves making judgments about threats, vulnerabilities and consequences, which are discussed below.
- A ***threat*** is a person or group of people, object or activity with the potential to cause harm to, for example, individuals, etc. In the AI system context, this includes developers and deployers, as well as past, present and future AI system activities. *Threat* is described above as one of the

factors related to risk, and typically it serves as an essential starting point in developing an understanding of AI system risk. For this reason, having an understanding of the environment in which AI systems are deployed and/or used is important in order to carry out an AI system risk assessment. In some instances, certain types of threat assessments might serve as a precursor for a AI system risk assessment.

- The concept of ***vulnerabilities*** as used in risk assessment comprises those things that can be exploited by the *threat* or that may support or facilitate its activities. In the AI System risk assessment context, looking at vulnerabilities as distinct from threat means focussing on, for example, the factors that represent weaknesses in AML/CFT systems or controls or certain features of a country. They may also include the features of a particular sector, a financial product or type of service that make them attractive for ML or TF purposes.
- ***Consequence*** refers to the impact or harm that an AI system may cause and includes its effect individuals' health, safety, fundamental rights and democracy. The consequences of an AI system may be short or long term in nature and also relate to populations, specific communities, the business environment, or national or international interests, as well as the reputation and attractiveness of a firm. As stated above, ideally a risk assessment involves making judgments about threats, vulnerabilities and consequences. Given the challenges in determining or estimating the consequences of AI systems, it is accepted that incorporating consequence into risk assessments may not involve particularly sophisticated approaches, and that countries may instead opt to focus primarily on achieving a comprehensive understanding of their threats and vulnerabilities. The key is that the risk assessment adopts an approach that attempts to distinguish the extent of different risks to assist with prioritising mitigation efforts.

Users of AI System Risk Assessments

The form, scope and nature of AI system risk assessments should ultimately meet the needs of its users – whether these are senior management, regulators, investors, users, etc. The number and diversity of users of an assessment varies according to the purpose for which it is carried out; however, typical users of risk assessments might include:

- Regulators and supervisors.
- Senior management for which the firm-level AI system risk assessment is a critical source in informing their risk-based obligations.
- International stakeholders.
- The general public, as well as academia, specified individuals, etc.

General Principles for Firm AI System Assessments

The general principles set out below could be considered when a firm intends to conduct any kind of AI system risk assessment. These include considerations on the purpose and scope of the assessment as well as the process through which an assessment will be conducted; the stages of a risk assessment, the participants, users and other parties involved; the information which may be used, and the final outcome of the assessment process.

The nature, methodology, participants, and information required for an assessment depend on the purpose and scope of the assessment. There is no single or universal methodology for conducting an AI system risk assessment. Therefore, this guidance does not advocate the use of any particular methodology or process. This guidance is aimed to provide a generic description of the risk assessment process as it might be applied to looking at risk associated with AI systems and considerations and practical tools for firms consider when undertaking AI system risk assessment.

Clear agreement on purpose

Before starting any kind of AI system risk assessment, all parties involved, including those who will conduct the assessment and, as appropriate, the eventual end users should be in agreement on the purpose and scope of the assessment. Expectations should also be set as to how the results relate to the understanding of firm-level risks. Generally, a AI system risk assessment is intended to help a firm to identify, assess and ultimately understand the AI system risks it faces. A firm may set out more concrete goals for a particular risk assessment however, such as informing the development of policy or the deployment of resources by employees and other stakeholders. Understanding the scale and impact of identified risks can also assist in determining the appropriate level and nature of AI system controls applied to a particular product or sector. Given the diversity of potential users and possible diverging expectations, it is essential at the outset that there be clarity about why an assessment is to be conducted, the questions it should answer, the criteria that will be used to answer those questions and the possible decisions that the assessment will feed into.

AI system risk assessments may be tied to strategic planning and linked to specific actions or decisions. For example, a firm AI system risk assessment serves as input to a firm AI system strategy or policy as part of the firm's international or domestic AI system co-ordination process. The purposes of the assessment will also vary according to the needs of the users. The purpose and scope of the assessment may also determine the methodology that is to be used.

Comprehensiveness of assessment

Regardless of the approach adopted, firms are advised to ensure that their assessment of AI system risk is comprehensive enough to provide an overall picture of the firm's AI system risks across the EU AI Act. Ideally, this picture should include sufficient breadth and depth about potential threats and vulnerabilities and their consequences to address the purpose and scope of the assessment. The range of threats and vulnerabilities relevant for any particular assessment will thus vary according to the scope of the assessment (department, divisional, etc.); however, the firm will need to ensure that all relevant risks are taken into account when the results from different types of assessments are combined to derive firm-level AI system risks.

Where information gaps exist or difficulties in reaching conclusions arise, it is useful if these can be recognised in the risk assessment and then become areas where more work is required in the future.

In addition, the uncertainty caused by the lack of information may itself raise the risk profile of the issue under consideration.

In seeking to develop a comprehensive picture, those in charge of the AI system risk assessment need to identify and acknowledge these limitations as they make a determination of the risks that can be assessed. Future risk assessments may be able to seek new or alternative sources of information that will permit assessment of areas that could not be adequately or fully assessed in an earlier work.

Need for high-level commitment to the AI System risk assessment process

Before conducting an AI system risk assessment, it is essential that there be the business will to carry out this work and ensure that the objectives of the assessment can be achieved. This business will may be demonstrated in a clear commitment from high-level business officials to the AI system risk assessment exercise. These officials will need to recognise, understand and acknowledge any AI system risks that exist within the firm and how these risks may be distinct from larger AI system-related threats. Situations where business officials purposely fail to identify AI system risks in their firm (or they deliberately determine certain risks as low level) because they believe that acknowledgement of a higher risk level may damage their reputation or may have a negative effect on investment within the firm need to be avoided.

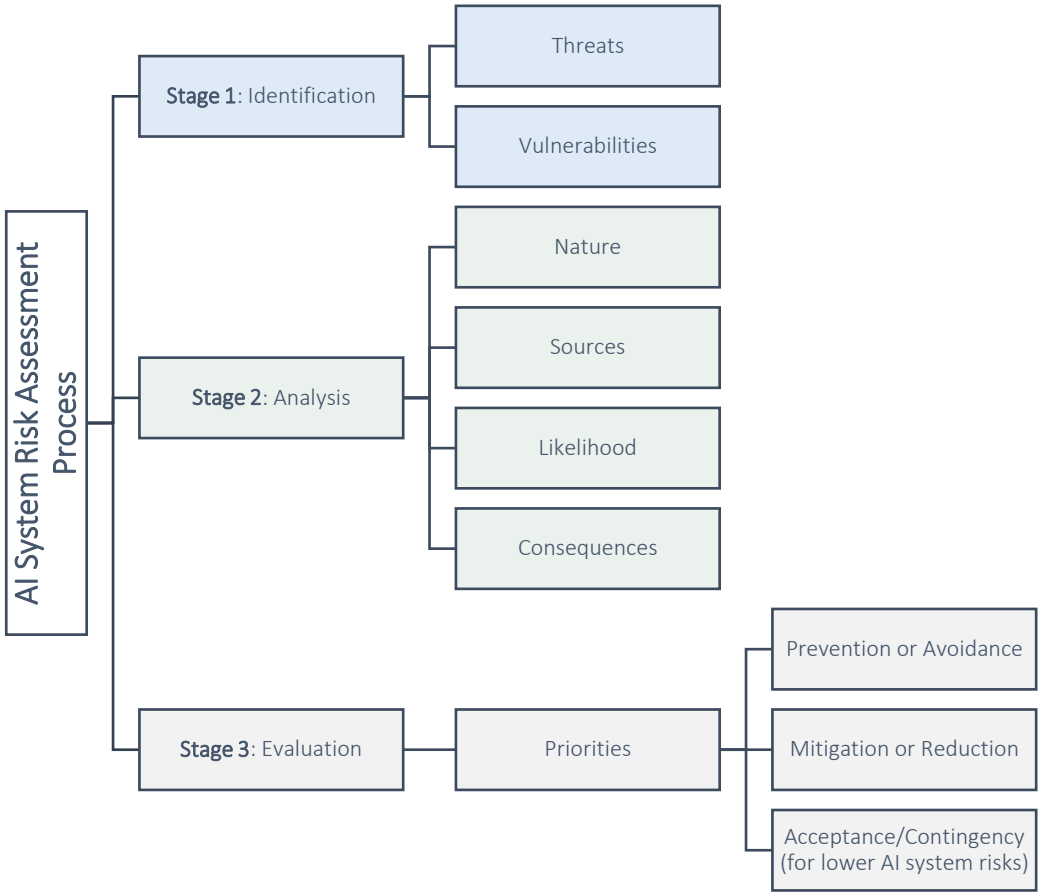
Appropriate judgment and balance are therefore important in the conduct of the firm-level AI system risk assessment process to prevent the process from becoming unduly influenced by or subordinate to a particular approach, agenda, resource injection, or lobbying by a specific stakeholder.

Stages of AI System Risk Assessment

The process of risk assessment can be divided into a series of activities or stages: ***identification***, ***analysis***, and ***evaluation***. The three stages are briefly described in this section. For completeness all three stages are described; however, this guidance focuses mainly on the first two. **Figure 1** below provides an overview of the AI system risk assessment process.

- In general terms, the process of ***identification*** in the context of an AI system risk assessment starts by developing an initial list of potential risks or risk factors firm face when deploying AI systems. These will be drawn from known or suspected threats or vulnerabilities. Ideally at this stage, the identification process should attempt to be comprehensive; however, it should also be dynamic in the sense that new or previously undetected risks identified may also be considered at any stage in the process.
- ***Analysis*** lies at the heart of the AI system risk assessment process. It involves consideration of the nature, sources, likelihood and consequences of the identified risks or risk factors. Ultimately, the aim of this stage is to gain a holistic understanding of each of the risks – as a combination of threat, vulnerability and consequence in order to work toward assigning some sort of relative value or importance to them. Risk analysis can be undertaken with varying degrees of detail, depending on the type of risk and the purpose of the risk assessment, as well as based on the information, data and resources available.
- ***Evaluation*** in the context of the AI system risk assessment process involves taking the risks analysed during the previous stage to determine priorities for addressing them, taking into account the purpose established at the beginning of the assessment process. These priorities can contribute to development of a strategy for their mitigation.

Figure 1: **Overview of the AI System Risk Assessment Process**



First Stage: Identification

After establishing the purpose and scope for the risk assessment exercise, a first step is to identify risks to be analysed. Given that AI system risks – as stated earlier in this guidance – are a combination of threats, vulnerabilities and consequences, a good foundation for the identification process is to begin by compiling a list of the major known or suspected threats and vulnerabilities that exist based on primary AI systems used or deployed. The identified AI system threats or vulnerabilities should of course relate to the purpose and scope of the assessment and this will also influence whether they are more micro or macro in focus.

At this initial stage, the list may be broad or specific, be based on actual or known typologies, or drawn from a more generic list of types of cases or schemes or circumstances involved in the AI system processes. For AI system threats, the development of a list may be facilitated by having access to, for example, typologies reports, as well as the collective knowledge of senior management. Formulating a list of the firm's major AI system vulnerabilities will typically be informed by reports by supervisors and about vulnerabilities in the sector, and the collective knowledge of the individuals involved in AI systems, particularly regarding the existence and effectiveness of any general mitigants or controls that help combat AI system risks. The exercise of establishing this first list of threats and vulnerabilities should consider the full process of AI systems, including any applicable contexts. Thus, discussion of AI system threats will probably need involvement of appropriate experts who contribute to compiling this initial list of the main or common AI system threats and vulnerabilities.

AI system risks exist when AI system threats exploit AI system related vulnerabilities. Thus after compiling a list of AI system threats and vulnerabilities, the next focus is for those involved in the process to think about how these interact and articulate a list of risks the firm faces when combating

AI system risks. It should be stressed that something identified on the list at this stage is not automatically classified as having higher (or lower) risk – it has simply been identified as sufficiently relevant to go into mix of risks to be analysed.

There are different approaches that may be used at the identification stage. One is based on identifying risk events, which involves starting from specific examples of AI system events – which may be macro or micro in nature. Under this approach the participants identify the main risk scenarios to analyse. Some examples of specific AI system risk events (derived from the threats, vulnerabilities and consequences) identified at this stage might include the following:

- Individuals or groups intentionally exploiting AI systems to perpetuate biases.
- Weaknesses in oversight allowing biased algorithms to go unchecked.
- Discriminatory outcomes affecting individuals' fundamental rights.
- Deliberate efforts to keep AI decision-making processes opaque.
- Insufficient regulatory measures to enforce transparency.
- Erosion of trust due to an inability to challenge or understand AI decisions.
- Intentional misuse of AI systems for invasive surveillance or unauthorized data access.
- Insufficient legal safeguards to protect personal data from AI processing.
- Negligence in considering environmental consequences in AI development.
- Lack of regulations addressing the energy-intensive nature of AI models.
- Increased carbon footprints contributing to environmental degradation.
- Rapid deployment of AI without thorough ethical considerations.
- Lack of foresight on potential negative consequences.
- Unforeseen negative impacts compromising safety or societal well-being.
- Intentional use of AI in ethically questionable scenarios.
- Absence of clear ethical guidelines or frameworks for AI development.
- Challenges to fundamental rights and democratic values.

Another approach that may be used starts from a macro-level and tends to focus more on circumstances. Under this approach a list of risk factors (relating to threats and vulnerabilities, see Annexes I and II for some examples of risk factors) is identified for analysis. The list can be expanded or narrowed down depending on the scope of the AI system assessment.

Irrespective of which approach is used for identification, those involved in the process must keep an open mind to ensure that all relevant risks or risk factors are identified so as to avoid inadvertently overlooking key issues that contribute to the firm's AI system risk. The actual processes used to identify the initial list of risks will vary. Some countries may utilise more formal techniques such as surveys and quasi-statistical analysis of past events or circumstances while others may carry out a brainstorming exercise among appropriate experts to produce a list or perhaps a tree diagram of related events or circumstances. Once an initial list of risks is identified, the assessment process can proceed to the next stage.

Second Stage: Analysis

Analysis lies at the heart of the AI system risk assessment process. It is through analysis that the process moves from a mere description of the AI system risks facing a firm– akin to a situation report – to fuller understanding of the nature, extent and possible impact of those AI system risks. As indicated in the introduction, risk can be thought of as a function of threat, vulnerability and consequence. The goal of this step is therefore to analyse the identified risks in order to understand their nature, sources,

likelihood and consequences in order to assign some sort of relative value or importance to each of the risks.

Understanding the Consequences Associated with AI Systems

In the process of analysing AI system risks, it is crucial to have a general understanding of why AI system risks arise. It is equally important to understand the consequences associated with the activity described above. This will assist in reaching conclusions about the relative importance of each identified risk. The consequences of AI system deployments are seen at the firm-level. From a firm perspective, one of the main consequences of AI systems is that it has a negative effect on the transparency, good governance and the accountability of firms. AI systems also cause damage to a firm's reputation and has a direct and indirect impact its commercial viability. Box 1 sets out examples of consequences of AI systems, to assist those carrying out AI system risk assessments to reach conclusions about the relative importance of each identified risk.

A particular challenge especially when using more formal techniques is that AI system risks are inherently difficult to describe or measure in quantifiable or numerical terms. It is therefore important to remember that risk as we have discussed it in this guidance is a combination of threats, vulnerabilities along with consequences. If the level of risk of the individual risks can be examined according to their consequences or impact and the likelihood of their materialising, then a rough estimate of risk level may be obtained. A very simple matrix as applied to a specific risk might be as shown in Figure 2.

Figure 2: Examples of Consequences of AI Systems (non-exhaustive)

Bias and Discrimination:

Potential violation of individuals' fundamental rights due to discriminatory outcomes in areas like hiring, lending, and law enforcement.

Lack of Transparency:

Undermining democratic principles by limiting accountability and hindering the ability to challenge or understand decisions made by AI systems.

Privacy Violations:

Infringements on individual privacy rights through the processing of vast amounts of personal data, leading to concerns about invasive surveillance, data breaches, and unauthorized use of personal information.

Environmental Impact:

Contribution to increased carbon footprints and environmental degradation due to the energy-intensive nature of training and running large AI models.

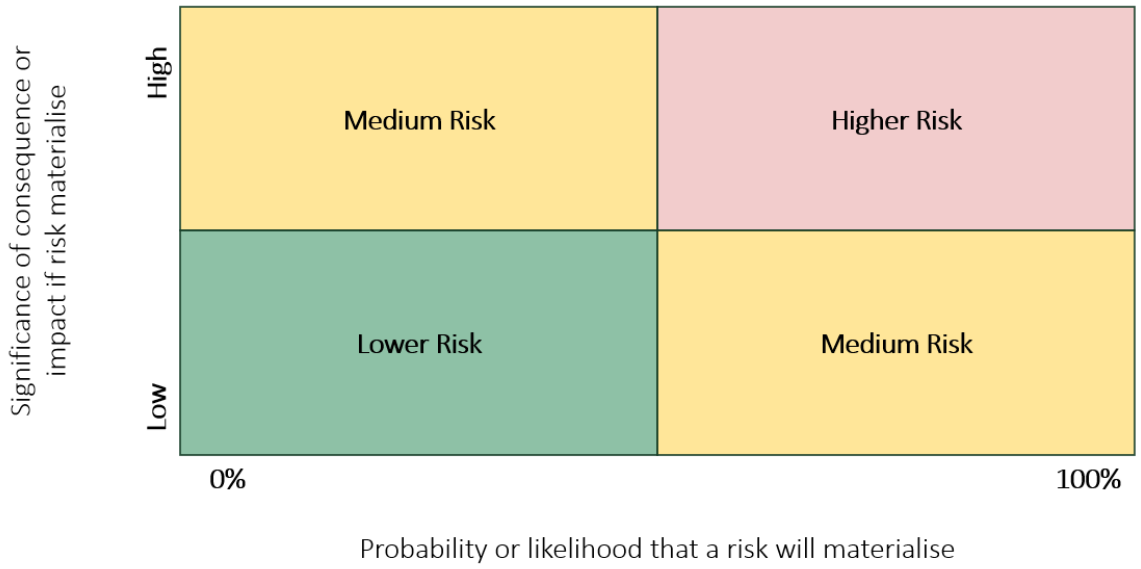
Unintended Consequences:

Production of unintended and potentially negative consequences due to unforeseen interactions or adaptations of AI systems to changing environments, compromising safety and societal well-being.

Ethical Concerns:

Challenges to fundamental rights and democratic values arising from ethical considerations, such as the use of AI in autonomous weapons or mass surveillance.

Figure 3: **Examples of a Risk Analysis Matrix**

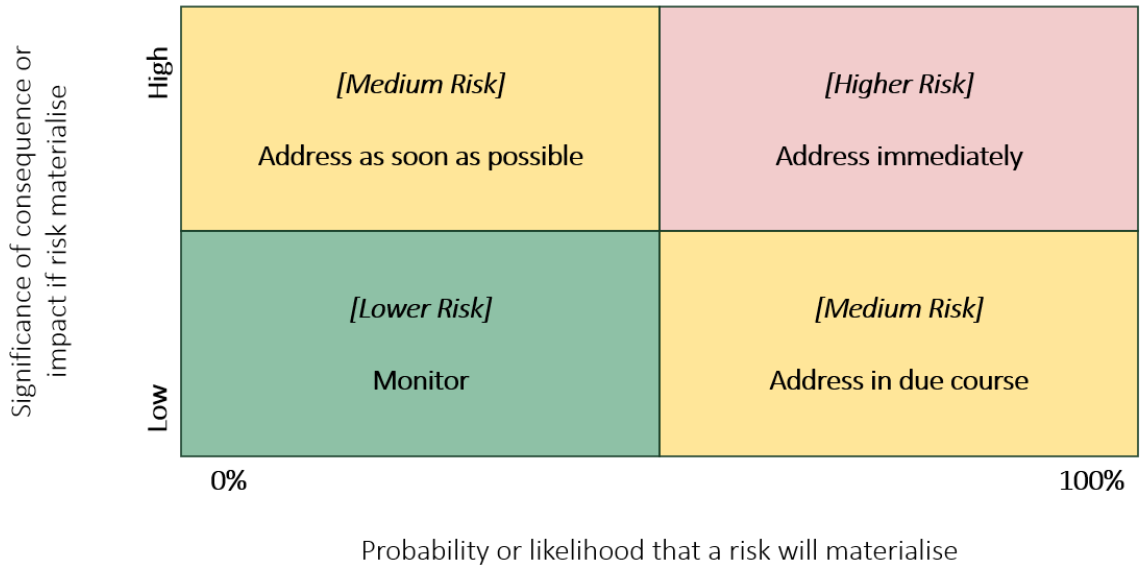


Third Stage: Evaluation

The last stage of risk assessment is evaluation. It involves taking the results found during the analysis process to determine priorities for addressing the risks, taking into account the purpose established at the beginning of the assessment process. These priorities can contribute to development of a strategy for their mitigation. As indicated in the introduction, this guidance does not attempt to provide a full explanation of this step of the process. For the sake of completeness however, some general details are set out here.

Depending on the source, there are a number of methods for addressing (or "controlling") risk, including prevention (or avoidance), mitigation (or reduction), acceptance or contingency planning. In the context of AI system risk and the risk-based approach, the most relevant of these methods are prevention (e.g., prohibiting certain AI systems) and risk mitigation (or reduction). The role of evaluating levels of AI system risk therefore normally leads to the development of a strategy for addressing the risks. Working from the example in the last section, the evaluation of risk levels for each of the analysed risks could result in courses of action as illustrated in Figure 4, which is provided as a simple example of how the evaluation process might proceed at this stage:

Figure 4: Examples of a Risk Evaluation Matrix



According to this example, higher levels of risk might require more immediate action to mitigate it; lower levels of risk might require lesser action or some other response (the example here indicates monitoring). Alternatively, higher levels of risk may indicate systemic or deeply entrenched risks which require a broader response over time. By their nature, such responses generally require consultation (within business functions, among others), and the implementation of measures, all of which can take time. The example shown here has been kept deliberately simple in order to clearly show the range of decisions that might be appropriate in addressing different levels of risk. A comprehensive AI system risk assessment process carried out at the firm level might use a more detailed matrix in order to encompass a wider range of potential actions. Also note that, other types of risk matrices than the examples given above or a list ranking of the risks may also work, but the basic principles of the concept of risk as discussed in this paper should be applied.

The prioritisation of AI system risks at the evaluation stage will assist in the challenge of allocating scarce resources to fund EU AI Act compliance programmes and other business efforts.

In the budgeting process, it is important to identify and prioritise issues that require attention. The evaluation process helps the firms make decisions about how best to utilise resources and set priorities for business units.

From an AI system context, firms should implement necessary measures (for example, the EU AI Act requirements) and allocate appropriate resources to mitigate the AI system risks which they have identified. In fact, the risk-based approach allows firms to develop a more flexible set of measures in order to target their resources more effectively, including by applying preventive measures flexibly across the firm. Based on the risks identified, measures should address how best to prevent the risk of harm arising from AI systems. Measures to mitigate AI system risk should also address the ways in which these actors can better detect and report this activity. From an operational perspective, measures should be in place to better detect, disrupt and managed AI system risks.

Outcome of Risk Assessments

The actual results of a risk assessment can take different forms. For firms that are ultimately the main users of the assessment, there is often an expectation that some form of a written report will be produced, although this is not strictly speaking a requirement of Article 9 of the EU AI Act. If the assessment will be presented in report form, decisions on how it will be organised – along with the level of detail – are most usefully made early on the risk assessment process and normally relate directly to the purpose and scope of the assessment. For example, a AI system risk assessment with end-users or other operational services as the primary users might discuss risks according to the threats that were the starting point of the assessment. For a report whose primary audience consists of regulators, a discussion of the AI system risks grouped according to vulnerability (AI system, etc.) might be most useful.

Regardless of the form and presentation of the AI system risk assessment, it should ultimately allow firms to make a judgment on the levels of the risks and priorities for mitigating those risks. The response can then be made commensurate to the nature and level of the risks identified. It is therefore advisable that the risk assessment contain sufficient information about the source, nature, and extent of each

risk to help indicate appropriate measures to mitigate the risk. Thus, the results of firm AI system risk assessments can provide valuable input in the formulation or calibration of firm-level AI system policies and action plans. These decisions may ultimately affect a number of business functions and how they carry out their responsibilities (e.g., how business activities are performed). The results of AI system risk assessments may also help inform planning for technical assistance on AI system matters by a broad range of stakeholders.

Dissemination of Assessments Outcome

Once completed, firms will have to consider how broadly the results of the risk assessment are to be disseminated amongst the various stakeholders.

Some AI system risk assessments may be considered to contain too much sensitive information to disclose publicly or that they may draw too much attention to the shortcomings in the AI system infrastructure of a firm. Furthermore, some of the information shared during the course of the assessment could be subject to confidentiality requirements. Nonetheless, appropriate information from assessments should be made available to assist it in addressing the current AI system risks and new and emerging threats. In certain firms, committees or working groups with vetted representatives have been created to share and discuss risk assessment information. More generally, it may be helpful to share information – at a minimum – on the main factors considered and the conclusions of the risk assessment process. Where the sensitive nature of the information prevents the broad distribution of the full results from the risk assessment report, consideration can be given to circulating sanitised information or summaries, or at least providing information on the methodology used, the findings and the conclusions. This approach could, for example, apply to information provided to assessors in the context of an AI system assessment.

A particular objective of a AI system risk assessment could be to provide information to stakeholders in order to enhance the general understanding of firm AI system initiatives. A typical output of a firm AI system risk assessment is generally a document. One challenge to overcome is that some information within the firm assessment may be derived from classified or sensitive sources. As such, some firms may produce a non-classified version.

Annex I. AI System Risks

Below are key risks associated with AI systems derived from Recital 43 & Article 14 of the EU AI Act.

Bias and Discrimination:

Bias and discrimination in AI systems refer to the unfair and prejudiced treatment of individuals or groups based on certain characteristics, such as race, gender, or socioeconomic status. This arises when the AI system's outputs or decisions reflect and perpetuate existing biases present in the training data.

Safety Concerns:

Safety concerns in AI systems involve the potential risks of harm or injury to individuals or damage to property resulting from the malfunction, misbehavior, or unintended actions of AI-driven technologies. This includes the safety of users, bystanders, and the broader environment.

Lack of Transparency:

The lack of transparency in AI systems refers to the opacity in decision-making processes, making it challenging for stakeholders, including end-users and the public, to understand how AI systems arrive at specific decisions or recommendations. This lack of clarity can hinder accountability and trust.

Privacy Violations:

Privacy violations in AI systems occur when personal information is processed or used without proper consent, leading to breaches of individuals' privacy rights. This risk involves the unauthorized access, use, or sharing of sensitive personal data.

Job Displacement and Economic Inequality:

Job displacement and economic inequality in the context of AI systems entail the potential negative impact on employment opportunities and income distribution. Automation and AI-driven technologies may lead to job losses and contribute to disparities in wealth and economic well-being.

Environmental Impact:

Environmental impact in AI systems refers to the effects of the technology on the environment, particularly concerning energy consumption, resource depletion, and contributions to carbon footprints. The training and deployment of large AI models can have significant environmental consequences.

Security Threats:

Security threats in AI systems involve the vulnerabilities that malicious actors could exploit for purposes such as cyberattacks, disinformation campaigns, or unauthorized access. These threats pose risks to the security of individuals, organizations, and even democratic processes.

Unintended Consequences:

Unintended consequences in AI systems refer to unforeseen outcomes or impacts that arise from the deployment or use of AI technologies. These consequences may manifest as unexpected behaviors, risks, or effects on individuals, society, or the environment.

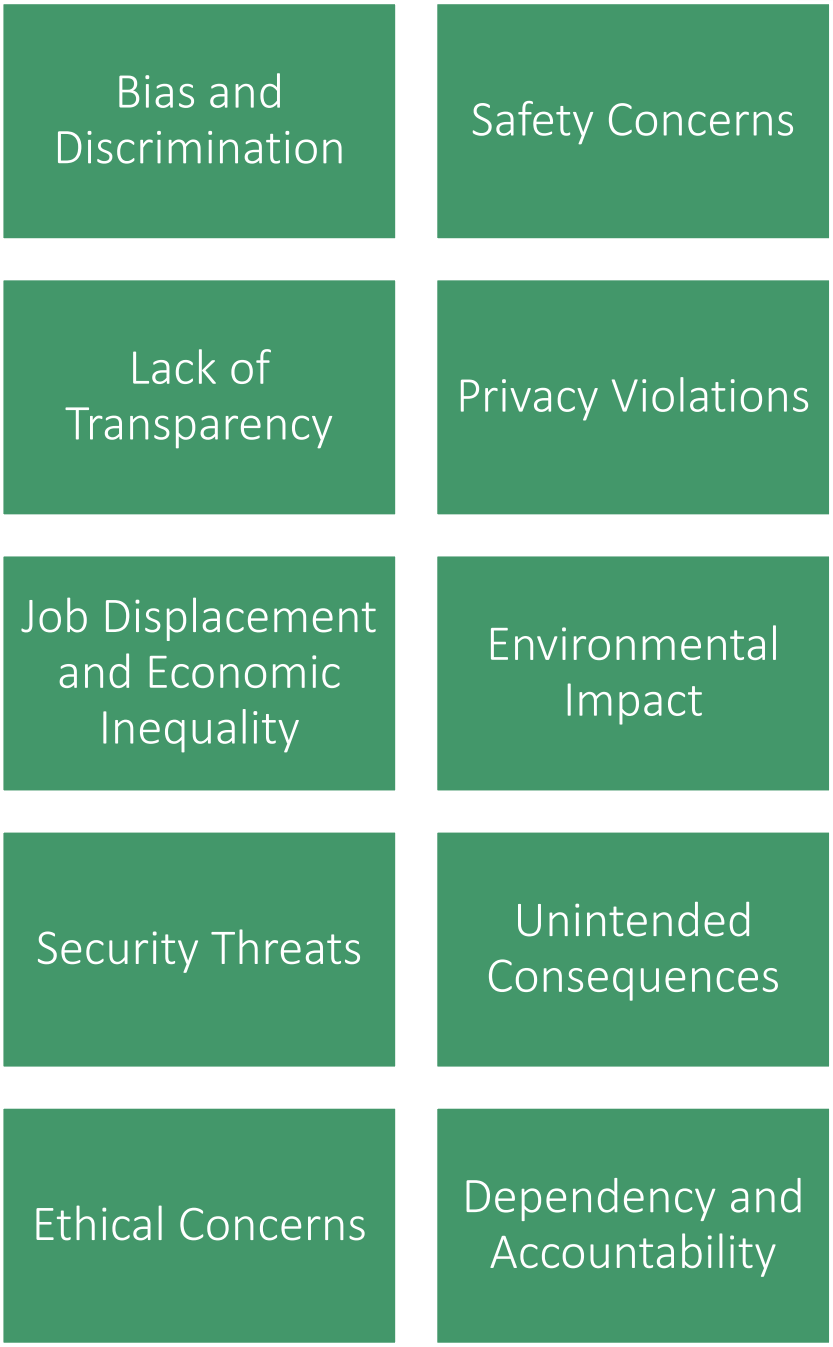
Ethical Concerns:

Ethical concerns in AI systems encompass issues related to the morality and principles governing the development and use of AI technologies. This includes considerations of fairness, accountability, transparency, and the responsible and ethical treatment of individuals and communities.

Dependency and Accountability:

Dependency and accountability in AI systems involve the overreliance on AI technologies without adequate mechanisms for accountability. This risk encompasses challenges in ensuring human control over AI systems, which could lead to a loss of accountability and potential negative consequences.

Figure 5: AI System Risks



Annex II. AI System Risk Factors Related to Vulnerabilities

- **Algorithmic Biases**: Inherent biases in algorithms leading to discriminatory outcomes.
- **Inadequate Security Protocols**: Weaknesses in security measures exposing AI systems to cyber threats.
- **Data Privacy Risks**: Lack of safeguards leading to unauthorized access and privacy violations.
- **Limited User Awareness**: Users unaware of potential risks, contributing to misuse.
- **Insufficient Model Explainability**: Lack of transparency in AI decision-making processes.
- **Biased Training Data**: Inclusion of biased data in training sets affecting AI performance.
- **Poor Governance Structures**: Absence of effective oversight leading to governance gaps.
- **Vendor Dependence**: Overreliance on single vendors creating dependency vulnerabilities.

- **Inadequate Compliance**: Failure to adhere to regulatory standards and compliance requirements.
- **Weak Authentication Mechanisms**: Vulnerability to unauthorized access due to weak authentication.
- **Unreliable Sensor Inputs**: Dependence on compromised or manipulated sensor data.
- **Overreliance on AI**: Blind dependence on AI without considering alternative approaches.
- **Lack of Regular Audits**: Infrequent assessments allowing vulnerabilities to persist.
- **Limited Interoperability**: Difficulty integrating AI systems with other technologies.
- **Incomplete Documentation**: Lack of comprehensive documentation on AI system architecture.
- **Ethical Decision-Making Gaps**: Lack of ethical frameworks guiding AI decision processes.
- **Insufficient Testing Protocols**: Limited testing procedures for identifying and addressing vulnerabilities.
- **Dependency Risks**: Lack of diversity in AI models leading to dependency on specific approaches.
- **Inadequate Redundancy**: Absence of backup systems, making AI vulnerable to failures.
- **Incompatible Standards**: Lack of industry-wide standards for AI system vulnerabilities.
- **Limited Cross-Domain Knowledge**: Developers lacking interdisciplinary knowledge for comprehensive risk assessment.
- **Cultural Resistance to Change**: Resistance to adopting necessary changes to mitigate vulnerabilities.
- **Adversarial Attacks**: Techniques used to manipulate AI systems by introducing manipulated inputs.
- **Incomplete Threat Intelligence Integration**: Failure to integrate timely threat intelligence into risk assessments.
- **Resource Allocation Gaps**: Inadequate allocation of resources for addressing vulnerabilities.
- **Insufficient Collaboration**: Limited collaboration hindering the identification and mitigation of vulnerabilities.
- **Poor Patching Procedures**: Delays or deficiencies in implementing patches to address known vulnerabilities.
- **Limited Training for Developers**: Developers lacking training on identifying and mitigating vulnerabilities.
- **Inadequate Public Accountability**: Mechanisms for holding developers and deployers accountable are insufficient.
- **Lack of Explainability Standards**: Absence of industry-wide standards for explaining AI decision-making processes.

Annex III. AI System Risk Factors Related to Threat

- **Malicious Use of AI**: Potential for AI systems to be employed for malicious purposes.
- **Cybersecurity Threats**: Risks of AI systems being exploited through cyber attacks.
- **Unintended Bias in Development**: Bias introduced during the development phase impacting AI fairness.
- **Manipulation by Adversarial Attacks**: Techniques to manipulate AI by introducing subtly crafted inputs.
- **Lack of Developer Accountability**: Developers engaging in activities that compromise system integrity.
- **Data Privacy Breaches**: Unauthorized access leading to the compromise of user data.

- **Inadequate System Authentication**: Weaknesses in system access controls leading to unauthorized use.
- **Insufficient Model Security**: Lack of protective measures for securing AI models from tampering.
- **Data Poisoning**: Introduction of malicious data during the training phase.
- **Inherent Algorithmic Biases**: Bias in algorithms leading to discriminatory outcomes.
- **Political Misuse**: Manipulation of AI systems for political purposes.
- **Developer Insider Threats**: Internal developers posing a threat to the system intentionally.
- **Lack of Adversarial Training**: Failure to train AI systems against adversarial inputs.
- **Supply Chain Vulnerabilities**: Risks associated with vulnerabilities in the AI supply chain.
- **Social Engineering Attacks**: Manipulating individuals to exploit AI system vulnerabilities.
- **Data Integrity Risks**: Risks associated with compromised integrity of training data.
- **Exposure to Inappropriate Content**: AI systems unintentionally generating or promoting inappropriate content.
- **Inadequate Regulatory Compliance**: Failing to comply with regulatory requirements and standards.
- **Geopolitical Threats**: Risks associated with geopolitical tensions impacting AI systems.
- **Lack of Explainability in Decision-Making**: Inability to understand and explain AI decision processes.
- **Inadequate System Monitoring**: Lack of real-time monitoring for detecting anomalies.
- **Public Perception Risks**: Negative public perception leading to rejection or misuse.
- **Intellectual Property Theft**: Risks of theft or unauthorized use of AI-related intellectual property.
- **Social Unrest Amplification**: AI contributing to the amplification of social unrest.
- **Competitive Espionage**: Rivals exploiting vulnerabilities in AI for competitive advantage.
- **Inadequate User Education**: Lack of awareness among users contributing to misuse.
- **Human-in-the-Loop Risks**: Risks associated with human involvement in AI decision processes.
- **Cross-Border Threats**: Threats originating from different legal jurisdictions.
- **Erosion of Public Trust**: Actions leading to a decline in public trust in AI systems.
- **Inadequate AI Impact Assessments**: Failure to assess the potential impact of AI on society comprehensively.