sciendo

Ethics & Bioethics (in Central Europe), 2024, 14 (1–2), 80–95 DOI:10.2478/ebce-2024-0007

Ethical considerations in Risk management of autonomous and intelligent systems

Anetta Jedličková[1]

Abstract

The rapid development of Artificial Intelligence (AI) has raised concerns regarding the potential risks it may pose to humans, society, and the environment. Recent advancements have intensified these concerns, emphasizing the need for a deeper understanding of the technical, societal, and ethical aspects that could lead to adverse or harmful failures in decisions made by autonomous and intelligent systems (AIS). This paper aims to examine the ethical dimensions of risk management in AIS. Its objective is to highlight the significance of ethical considerations in mitigating risks associated with the development, deployment, and use of AIS. The paper provides an overview of various types of AI risks and risk management procedures aimed at mitigating the negative impacts of those risks. We employ a comprehensive risk management approach that combines technical expertise with ethical analysis to ensure alignment with human values and societal objectives. Through the analysis of AI risks and risk management procedures, we advocate for establishing effective mechanisms for ethical oversight and legal control to promote ethical and trustworthy AIS. The findings reveal key risks associated with transparency, accountability, privacy infringement, algorithmic bias, and unintended consequences. To address these challenges, we consider integrating ethical principles into risk management practices, transparent risk communication, continuous engagement with all stakeholders, establishing robust accountability mechanisms, and regular ethical oversight as imperative in ethically designing and operating AI systems. Given the diminished effectiveness of internal audits compared to external audits, we also recommend the implementation of regular monitoring mechanisms through independent external audits when evaluating risk management practices.

Keywords: AI audits, AI ethics, AI risks, AI risk management, autonomous and intelligent systems, ethical risk assessment

Introduction

In recent years, significant efforts have been made to develop autonomous and intelligent systems (AIS) across various industries, including the economy, business, transportation, and healthcare sectors, impacting all areas of human activities. Ethical concerns regarding the impact of AIS have prompted increasingly important discussions concerning the security of algorithmic decision-making. It is necessary to carefully examine the impact of AIS on human rights, as the autonomy of AI applications and the inherent uncertainty in their operation can potentially inflict societal harm on diverse individuals and social groups. To avert harm caused by AI applications, developers of AI algorithms must take full responsibility for respecting the human rights of all users. They should ensure continuous human control and oversight over automated algorithmic decision-making. Moreover, relevant stakeholders should develop procedures and processes to assess the security of AIS and appropriately allocate responsibilities, rights, and duties.

The use of AI applications introduces potential security risks, underscoring the importance of ensuring that AI technologies are used in an ethical, transparent, and responsible manner. It is important to guarantee that relevant stakeholders, including users, engage with trusted data and possess knowledge of secure practices in their utilization. Effective risk management processes and procedures can support the practical phases of risk governance, covering the assessment of all risks associated with AI. These processes provide efficient methods to enhance the efficacy of control and oversight measures, focusing on pertinent actions to mitigate risks that may negatively affect human rights, dignity, privacy, health, safety, or security. AI risk management requires a systematic approach to monitoring, governing, and overseeing safety, entailing processes for identifying, evaluating, and managing risks associated with AI

[1] Charles University Prague (Czech Republic); Anetta.Jedlickova@fhs.cuni.cz; ORCID: 0000-0003-1239-4046

technologies. This involves the systematic application of policies and procedures to activities involving the assessment, monitoring, communication, reporting, and treatment of risks related to AIS products, services, and data processing. Ethical considerations in risk management help identify and mitigate the adverse consequences of governance gaps that deviate from human values and societal objectives by determining ethical requirements.

The state of the art in AI risk management

The National Institute of Standards and Technology (NIST), a leading institution in the development of AI standards, released the initial version of the *Artificial Intelligence Risk Management Framework* in January 2023. This framework, designed for voluntary use, aims to enhance the integration of trustworthiness into the design, development, use, and assessment of AI systems. It defines AI risk management as a means to minimize potential negative impacts on human rights and freedoms, while maximizing positive impacts (NIST, 2023).

In the European Union (EU), the use of AIS will be regulated by the *Artificial Intelligence Act* (AI Act), set to become the world's first comprehensive AI law. Negotiations commenced among EU member states in June 2023 to finalize the AI Act, following amendments adopted by the European Parliament to the initial proposal for harmonized AI rules, published in April 2021. In December 2023, negotiators from the EU Council presidency and the European Parliament reached a provisional agreement on the proposed AI Act, and on March 13, 2024, the EU Parliament officially approved the act. This regulation aims to ensure AI systems used in the EU are safe, respect fundamental rights and values, and protect democracy, rules of law, and environmental sustainability against high-risk AI while fostering innovation. In terms of governance and compliance, the AI Act establishes an AI Office within the European Commission to oversee the most advanced AI models. It introduces new obligations for providers and users based on the level of potential risks and the level of their impacts, which need to be assessed during risk management procedures. This involves a continuous iterative process throughout an AI system's lifecycle. To maintain effectiveness, regular reviews and corresponding updates of the risk management process are necessary, along with documenting significant decisions and actions taken. The list of high-risk areas should undergo ongoing review during regular risk assessments (AI Act, 2023a; AI Act, 2023b).

The United Nations Educational, Scientific and Cultural Organization (UNESCO) accentuates the increasing importance of global endeavors to ensure the ethical development of science and technology. UNESCO has established international standards through the *Recommendation on the Ethics of Artificial Intelligence* to maximize the benefits of scientific and technological advancements while minimizing associated risks. These standards advocate for the establishment of risk assessment procedures and ethical impact assessment frameworks to identify and evaluate the benefits, concerns, and risks of AI systems. This includes implementing measures for risk prevention, mitigation, and monitoring, as well as assessing impacts on human rights and fundamental freedoms, particularly for marginalized and at-risk groups or persons in vulnerable situations (UNESCO, 2022).

While various frameworks, regulations, and standards concerning AI ethics emphasize the significance of the risk management process in ethically evaluating trustworthy AI systems, they lack practical guidance on executing risk assessments. A comprehensive theoretical guideline for risk management processes and the establishment of effective risk management frameworks is provided by the *ISO 31000 (Risk Management – Guidelines)*. Since the ISO 31000 is not industry or sector-specific, its procedures can be customized to the needs of any organization and applied to diverse activities and risks faced by organizations at all levels (ISO 31000, 2018). The *IEEE Standard for Software Life Cycle Processes – Risk Management*, developed by the Institute of Electrical and Electronics Engineers (IEEE) standards boards, provides detailed requirements related to AI risks to facilitate the acquisition, supply,

development, operation, and maintenance of software products and services (IEEE, 2001). Additionally, the Organisation for Economic Co-operation and Development (OECD) has offered a comprehensive overview of integrating risk-management frameworks throughout the entire AI system lifecycle to promote trustworthy AI. This is outlined in the recent OECD report entitled *Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI)* (OECD, 2023).

While many authors discuss ethical issues associated with the risks of AI systems, there is no clear strategy in the literature on implementing AI risk management into practice. Some authors examined cyber risk management (Paté-Cornel et al., 2018; Eling, McShane & Nguyen, 2021), with the latter reviewing cybersecurity research history, assessing the cyber risk management process, highlighting gaps, and determining research directions. The authors emphasize the need for interdisciplinary collaboration in cyber risk management (Eling, McShane & Nguyen, 2021). Others explored a risk-based approach (Fraser & Bello y Villarino, 2021; Chamberlain, 2023; Schuett, 2023) with the latter analyzing key risk management provisions in the AI Act and proposing ways in which the risk management provision can be amended. Additionally, Macrae (2022) proposed a theoretical framework to identify and analyze sociotechnical sources of AI risks. Based on analysis of the fatal self-driving Uber accident, the framework identifies and characterizes five fundamental interconnected sources of sociotechnical risks in AI systems such as structural, organizational, technological, epistemic, and cultural.

This paper contributes to the discussion on AI risk management by providing an overview of implementing risk management processes in the ethical assessment of AIS, addressing various risks associated with AIS.

Ethical risk-based design process

The *IEEE Standard 7000™* outlines an ethical risk-based design process aimed at assisting engineers and technologists in integrating ethical considerations into their system designs to address risks associated with AI applications. By designing systems that incorporate individual and societal ethical values, such as transparency, accountability, privacy, fairness, security, safety, efficiency, and effectiveness, organizations can ensure that their systems meet both functional and ethical requirements. Embedding ethical values into the design process of an AIS and ensuring their adherence is paramount. This ethical risk-based design process aids in identifying and mitigating risks to ethical requirements. This process involves consulting with stakeholders to identify risks and associated contexts, estimating the likelihood of hazards or harm occurring, evaluating the nature and magnitude of consequences and general impacts of the assessed AIS, and determining appropriate risk treatments. The risk level for a given value can be ascertained by combining the likelihood of hazards or harm with their severity. Technical and organizational controls over the system and its components should be analyzed and specified, including control mechanisms for system functions that may impact values-based requirements. Regular verification of the effectiveness and acceptability of the implemented controls and risk reduction options is essential (IEEE, 2021).

Han et al. (2022) underscored the importance of understanding human values for advancing ethics and morality in AI. They emphasized that embedding human values into AI system functions is essential to ensure that the system makes evaluative choices aligned with human objectives and values, thus guaranteeing safe and effective implementation. Accurate knowledge of human values is pivotal for incorporating the correct value "codes" into AI algorithms. AI systems, characterized by autonomy, adaptability, and interactivity, evolve and acquire numerous features during operation. However, this evolution poses the risk of AI systems embodying values that diverge from human values. Therefore, continuous monitoring of the realized values of AIS is critical, along with undertaking redesign activities to address

any unintended and unforeseen consequences. Ongoing monitoring and redesign efforts help ensure that AI systems embody the values deemed important by society, whether established during the initial system design or recognized over time due to unintended outcomes. Sustaining the alignment of AIS with the right values requires persistent human oversight (Poel, 2020).

Types of AI risks

A risk, as defined in the *IEEE Standard for Software Life Cycle Processes – Risk Management*, represents the likelihood of an adverse event, hazard, threat, or negative situation occurring along with its undesirable consequences (IEEE, 2001). AI risks refer to the potential harms that may arise from the development, deployment, and utilization of AI systems, impacting individuals, social groups, organizations, systems, and the environment. These risks can stem from various sources, including the data used to train and test the AI system, the system itself (e.g., the algorithmic model), the manner and environment in which AI systems are employed, or their interaction with human beings. Different risk categories exist depending on the nature of the risk, such as technical, legal, organizational, financial, operational, business, strategic, reputational, safety, and others.

ForHumanity, a nonprofit organization focused on addressing ethics, bias, privacy, trust, and cybersecurity in AIS, has identified 13 key *risk categories* that warrant assessment. These categories include Privacy, Security, Safety, Bias, Governance, Ethics capability, Transparency, Explainability, Accountability, Accessibility, Diversity, Human agency, and Sustainability (Carrier & Narayanan, 2022).

To avoid exposure to adverse impacts, the risk management process should encompass the assessment of activities that could generate risks and should perform a thorough and systematic root cause analysis. As mentioned earlier, the AI Act adopts a risk-based approach, meaning that stricter regulations apply to higher-risk scenarios. It establishes distinct regulations for providers and users based on the varying levels of risks posed by AIS. Depending on the potential risk impacts of AI applications, the AI Act categorizes AI systems into those posing *unacceptable risk, high risk, and limited risk*. AI practices associated with unacceptable risk will be prohibited due to the potential for intolerable harm. High-risk AI systems will be permitted but subject to a set of requirements and obligations to obtain access to the EU market. AI systems presenting limited risk will be subject to limited transparency obligations (AI Act, 2023a; Madiega, 2023).

1. Unacceptable risks of AI systems

According to the provisions of the AI Act proposal, AI systems deemed to pose an unacceptable risk and threaten human safety or rights will face a ban. Such systems may include:

- AI systems employing manipulative subliminal techniques to carry out cognitive behavioral manipulation of individuals or specific vulnerable groups.
- AI systems utilized for social scoring purposes categorizing individuals based on behavior, socio-economic status, or personal characteristics.
- Real-time remote biometric identification systems deployed in publicly accessible spaces. The technical inaccuracies inherent in AI systems designed for remote biometric identification of individuals may produce biased results and pose discriminatory effects. Given the limited opportunities for subsequent checks or corrections in real-time operation, these systems pose increasing risk to individuals' rights and freedoms. Therefore, their use in publicly accessible spaces for law enforcement purposes should be prohibited. Additionally, AI systems used to analyze recorded footage from publicly accessible spaces through post-remote biometric identification systems should also be banned, except under pre-judicial authorization for use in law enforcement, strictly when necessary for targeted

searches related to specific serious criminal offenses that have already occurred, and only with a pre-judicial authorization (AI Act, 2023a).

2. High risks of AI systems

AI systems classified as high-risk AIS are those that have a detrimental impact on health, safety, or fundamental human rights. As outlined in the proposed AI Act, fundamental human rights encompass a broad spectrum of rights, including the right to human dignity, respect for private and family life, protection of personal data, freedom of expression and information, freedom of assembly and association, non-discrimination, the right to education, consumer protection, workers' rights, the rights of persons with disabilities, gender equality, intellectual property rights, the right to an effective remedy and a fair trial, the right of defense, presumption of innocence, and the right to good administration (AI Act, 2023a).

Additionally, AI systems serving as safety components of products, or those constituting products themselves falling under specific EU harmonization legislation, are deemed high-risk if the product undergoes the *ex-ante* conformity assessment procedure to ensure compliance with safety requirements. Examples of such products include toys, lifts, medical devices, and machines.

Furthermore, AI systems falling into the following categories are also classified as high-risk:

- *Biometric identification and categorization of individuals, including facial recognition technologies*: This category encompasses AI systems analyzing biometric or biometrics-based data to infer personal characteristics, including emotion recognition. New provisions in the AI Act mandate users of emotion recognition systems to inform individuals when they are subject to such systems (AI Act, 2023a; AI Act, 2023b).
- *Management and operation of critical infrastructure*: AI systems serving as safety components in managing road, rail, and air traffic, as well as supplying essential services like water, gas, heating, electricity, and critical digital infrastructure (AI Act, 2023a).
- *Education and vocational training*: AI systems used in determining access or influencing decisions related to admissions or assignments in educational and vocational training institutions (AI Act, 2023a).
- *Employment, worker management, and access to self-employment*: AI systems used in the recruitment and selection of individuals, job advertisement targeting, screening or filtering applications, and candidate evaluation in the course of interviews or tests. Additionally, this category includes AI systems used in decision-making regarding work-related contractual relationships, task allocation based on individual behavior or personal traits or characteristics, performance monitoring, and behavior evaluation (AI Act, 2023a).
- *Access to and enjoyment of essential private services, public services, and benefits*: AI systems used by or on behalf of public authorities to assess eligibility for public assistance benefits and services, including healthcare, housing, electricity, heating/cooling, and internet, as well as to grant, reduce, revoke, increase or reclaim such benefits and services (AI Act, 2023a).
- *Law enforcement*: AI systems, such as polygraphs, employed by or on behalf of law enforcement authorities to assess evidence reliability during criminal investigation or prosecution (AI Act, 2023a).
- *Migration, asylum, and border control management*: AI systems used by competent authorities to assess risks, including security risks, risks of irregular immigration, or health risks, posed by individuals entering or intending to enter EU Member States. This category includes AI systems used to verify the authenticity of travel documents and detect non-authentic documents by checking their security features (AI Act, 2023a).
- *Administration of justice and democratic processes*: AI systems assisting judicial or administrative bodies in researching, interpreting facts and law, applying the law, or in

alternative dispute resolution. However, AI tools should support rather than replace human-driven decision-making (AI Act, 2023a).

According to the AI Act, all high-risk AI systems must comply with a specific set of requirements, focusing on risk management, testing, technical robustness, data training and governance, transparency, human oversight, and cybersecurity (AI Act, 2023a). As emphasized by the European Parliament, high-risk AI technologies are expected to adhere to a set of principles that include safety, transparency, accountability, non-bias or non-discrimination, social responsibility, gender equality, rights to redress, environmental sustainability, privacy, and good governance. These principles should be upheld through impartial, objective, and external risk assessments conducted by national supervisory authorities. To ensure trustworthiness, high-risk artificial intelligence, robotics, and related technologies, including the software, algorithms, and data used or produced by such technologies should be developed, deployed, and used in a safe, transparent, and accountable manner. This entails incorporating safety features such as robustness, resilience, security, accuracy and error identification, explainability, interpretability, auditability, transparency, and identifiability. Moreover, AI technologies should be designed and developed to allow for the disabling of specific functionalities or a return to a previous state to restore safe operations in cases of non-compliance with the aforementioned features. In terms of transparency, public authorities should have access to the technology, data, and computing systems underlying such technologies when strictly necessary (European Parliament, 2020).

3. Limited risks of AI Systems

AI systems classified as presenting limited risk are subject to a limited set of transparency requirements. These requirements aim to provide users with the necessary information to enable informed decision-making regarding the utilization of a specific AI system. Ensuring users are aware when they interact with AI technology is paramount. Examples of systems falling under this category may include those generating or manipulating image, audio, or video content, as well as those involving human interaction, such as chatbots.

Additionally, per the provisions of the AI Act, AI systems posing low or minimal risk can be developed and deployed within the EU without the need to fulfill any supplementary legal obligations. Nonetheless, the act recommends that providers of such AI systems voluntarily adhere to the mandatory requirements applicable to high-risk AI systems (AI Act, 2023a).

Risk management

Risk management aims to proactively identify potential issues before they manifest allowing for the implementation of suitable measures to reduce or eliminate the likelihood of adverse impacts, and in the event of their occurrence, to mitigate the associated risks and impacts (IEEE, 2001). AI risk management activities include the processes of identification, assessment, and management of risks linked with AI technologies, addressing both technical and non-technical aspects. To execute such activities effectively, it is imperative to thoroughly understand the potential AI risks, possess the capabilities to analyze them properly, and develop strategies for their mitigation. AI risk management also entails establishing procedures and systems to ensure compliance with ethical and legal standards, alongside appropriate internal and external policies.

A risk management process, as defined by the IEEE standard, constitutes a continuous process involving the identification, analysis, treatment, and monitoring of risks throughout the lifecycle of AI products or services. It should adopt a systematic approach to address anticipated or occurring risks across the entirety of an AIS lifecycle (IEEE, 2001). Risk management processes should encompass assessments of all AI-related risks, offering effective methods to enhance process efficiency, and focusing on pertinent actions to mitigate risks that may

negatively impact human rights, dignity, privacy, health, safety, or security, as well as transparency, explainability, accountability, accessibility, or other ethical requirements.

Risks are assessed based on their potential negative impacts on humans, society, and the environment. As such, the main *types of risk impacts* are as follows:

- Impact on individuals/groups (e.g., reputation damage or identity compromise)
- Societal impact (e.g., disruption of democratic systems, influence on social policies)
- Environmental impact (e.g., climate damage)

Effective risk management is an indispensable aspect of decision-making and should be seamlessly integrated into an organization's structure, operations, and processes. It is essential to recognize that both the focus and methodology of risk management can evolve across different phases of the AI lifecycle. Organizations should define the extent and nature of risks they are willing to accept or mitigate. Specific criteria should be established to assess the significance of a risk, aligning with the organization's values, objectives, and resources. These criteria should undergo periodic review and adjustment as necessary.

The risk management process involves the systematic application of policies and procedures to *assess, monitor, communicate, report, and treat* risks associated with AIS products, services, and data processing (ISO 31000, 2018). Regular reviews and ongoing monitoring of all relevant activities affecting the validity, reliability, accuracy, interpretability, or performance of AIS outcomes are strongly recommended. All personnel within the organization should possess a comprehensive understanding of risks and remain informed about risk-related issues, risk management practices, and fundamental procedures. Risk managers bear the responsibility of identifying, analyzing, assessing, documenting, and reviewing risks, as well as implementing and overseeing appropriate risk management control measures based on the identified risk levels.

1. The risk management approach is primarily based on the *risk assessment process* which comprises three-step phases: *risk identification, risk analysis, and risk evaluation*.
    - *Risk identification* entails examining and understanding causal relationships and interactions among risk indicators. Crucial factors such as the potential for harm, the likelihood of harm occurrence, the feasibility of corrections, risk sources, threats, opportunities, vulnerabilities, capabilities, knowledge limitations, time-related factors, and others should be considered. It is essential to examine both the intended use and possible misuse of AIS, along with knowledge of any negative impacts and adverse incidents gleaned from AIS monitoring. A comprehensive list of identified risks, including risk categories, type of risk impacts, and proposed risk treatments should be precisely maintained.
    - *Risk analysis* focuses on evaluating the impact and potential adverse outcome of identified risk indicators on human subject protection, society, and the environment within the context of risk categories. Each adverse outcome should be linked to the relevant risk categories and impact types outlined above. This phase involves a detailed examination of uncertainties, sensitivity levels, the nature and extent of consequences, event and consequence likelihood, and the effectiveness of existing controls.
    - *Risk evaluation* involves assessing and determining the risk level based on the severity and likelihood classification. The outcomes of this assessment are then compared to the established risk criteria to determine whether additional actions are required. This phase is crucial in supporting decision-making processes. AI risks can be evaluated across different levels: governance and process levels focusing on risks potentially affecting value-based principles such as transparency and accountability, or technical levels addressing technical risks such as robustness and performance (ISO 31000, 2018).

2. *The risk monitoring process (including reviews and controls)* is focused on mechanisms identifying the need for any reassessment procedures. This process includes defining roles, responsibilities, and systematic safeguards to ensure compliance with standard operating procedures. The approach employed to mitigate risk to an acceptable level should be proportionate to the risk's significance. Periodic reviews of risk control measures are conducted to gauge the effectiveness and relevance of implemented risk and quality management activities. If deemed necessary, adjustments are made to the process design or implementation. Monitoring and review activities should be conducted in all stages of the risk management process (ISO 31000, 2018).
3. *The risk communication process* entails continuous notifications aimed at facilitating the (re)assessment and treatment of risks. It encompasses communicating pertinent information regarding risk and quality management activities to stakeholders involved in or affected by these initiatives. This practice enhances stakeholders' awareness and comprehension of risks, thereby facilitating ongoing risk oversight, regular risk reviews, and continuous improvement efforts. Additionally, the process involves consulting with independent external experts to bring diverse areas of expertise into the process (ISO 31000, 2018).
4. *The risk reporting process* entails determining the necessity and frequency of recording and reporting risks based on their risk levels to relevant stakeholders. This process also encompasses disseminating information about risk management activities and outcomes, including adverse incidents, across the organization. Stakeholders should be regularly informed about the risk and quality management approach applied in the AIS, significant deviations from the predefined risk/quality tolerance limits, and any remedial actions taken (ISO 31000, 2018).
5. *The risk treatment process* is pivotal in mitigating risk impacts. It involves employing techniques to prevent, mitigate, or eliminate risks, considering their likelihood and impact. This process focuses on implementing suitable measures to mitigate risks posed by AIS, minimizing adverse impacts attributable to these risks, and ameliorating negative consequences associated with specific AIS products or services. Initially, stakeholders need to determine whether a risk is acceptable. If not, a prioritization of identified risks for treatment should be established, and applicable risk treatment options formulated. Subsequently, the most appropriate risk treatment option is selected by weighing potential benefits against the costs, efforts, or disadvantages associated with the selected treatment implementation. Finally, appropriate actions are initiated to reduce risks to an acceptable level. The effectiveness of the treatment is assessed, and if the residual risk remains unacceptable, further treatment is taken. Continuous monitoring and review of the effectiveness of risk treatment implementation are essential, as risk treatments may produce unintended consequences. Moreover, risk treatment may introduce new risks that require risk management (ISO 31000, 2018).

   There are two complementary approaches to risk treatment:

   - *Process-related approach*, based on procedural, administrative, and governance mechanisms.
   - *Technical approach*, focusing on the technological specifications of an AI system.

   Conflicts and interactions may arise between procedural and technical measures. For instance, removing bias might cause a loss of accuracy, or increasing explainability might impact privacy. Trade-off analyses aim to optimize the balance for such cases in legal and ethical contexts (OECD, 2023).

Case studies, key insights, recommendations

The following case studies provide illustrative examples of risk management considerations within the realm of AI.

1. *Artificial Intelligence in Risk Management – KPMG*: This case study delves into the transformative impact of AI and Machine Learning techniques in risk management practices. It highlights how these technologies can enhance risk management efforts by improving forecasting accuracy, optimizing variable selection processes, and enabling richer data segmentation. The unique emphasis lies in exploring the transformative potential of AI within a specific industry while addressing ethical implications (KPMG, 2021).
2. *Derisking AI: Risk Management in AI Development – McKinsey*: This case study explores the impact of AI across various business operations, encompassing customer service, marketing, training, pricing, and security. It underscores the necessity of integrating risk management directly into AI initiatives. The unique emphasis is on the importance of constant controls and oversight throughout the development and deployment of AI (Baquero et al., 2020).
3. *The Case for AI Insurance – Harvard Business Review*: This case study introduces the innovative concept of AI insurance as a strategy to manage the risks associated with the integration of machine learning systems into business operations. The unique emphasis is on exploring innovative risk management approaches (Kumar, Shankar & Nagle, 2020).

In summary, these case studies provide valuable insights into AI risk management practices and emphasize the transformative potential of AI. While all three case studies discuss AI risk management, each case study offers a unique perspective and emphasizes different aspects of the topic. This diversity of viewpoints contributes to a more comprehensive understanding of AI risk management.

Several common threads connect the presented case studies:
1. *Emphasis on risk management*: All three case studies highlight the critical role of risk management in AI. Whether it is KPMG's application of AI in financial services, McKinsey's emphasis on derisking AI, or the Harvard Business Review's discourse on AI insurance, each case study accentuates the necessity of robust risk management strategies in AI.
2. *Dual nature of AI*: The case studies also underscore the dual nature of AI. While AI has the potential to revolutionize various industries, it also introduces new and varied risks. This dual nature of AI is a common theme across all three case studies.
3. *Call for innovative approaches*: Each case study emphasizes the need for innovative approaches to managing AI risks. KPMG discusses employing AI techniques in risk management, McKinsey focuses on integrating risk management directly into AI initiatives, and the Harvard Business Review advocates for AI insurance. These innovative approaches are necessary to manage the unique risks posed by AI.
4. *Role of stakeholders*: Lastly, all three case studies highlight the pivotal role of various stakeholders - from developers and users to regulators and insurers - in managing AI risks. This underscores the fact that managing AI risks is a shared responsibility.

In conclusion, while each case study focuses on a different aspect of AI risk management, they are all linked by their emphasis on the importance of risk management, the dual nature of AI, the need for novel approaches, and the involvement of stakeholders in managing AI risks. These common threads provide a comprehensive perspective on the challenges and opportunities in managing AI risks.

Key insights include:
1. *Ethical considerations and transparency*: AI can significantly enhance risk management, but it must be underpinned by ethical considerations and transparency.

2. *Constant oversight and employee empowerment*: Continuous oversight and empowering employees with the requisite knowledge and skills are crucial for managing AI risks effectively.
3. *Exploration of innovative solutions*: Innovative solutions such as AI insurance could serve as a safety net for organizations, but their implementation requires careful consideration due to their inherent complexities.

Derived from these insights, the following recommendations are proposed:
1. *Ethical guidelines and transparency*: Organizations should develop robust ethical guidelines for AI use in risk management. They should invest in explainable AI to make the decision-making process understandable to humans.
2. *Employee training and empowerment*: Investing in training programs to educate employees about AI risks and fostering a culture of responsibility and accountability are essential.
3. *Innovative solutions*: While exploring innovative solutions, organizations must address the complexities associated with their implementation.

It is imperative to explore methods of empowering employees to comprehend and manage the risks associated with AI. Additionally, we need to evaluate innovative approaches to managing AI risks, such as AI insurance, ensuring that the data utilized by these innovative risk management tools remain unbiased and that their predictions are transparent and explainable. Addressing these complex challenges is paramount for the successful implementation of innovative risk management approaches. Furthermore, a reformation of our existing approaches to AI risk management is essential. This transformation will likely involve a combination of ethical considerations, employee empowerment initiatives, and the adoption of innovative solutions.

Ethical risk assessment

Risk management also includes establishing processes to ensure compliance with ethical and legal standards. Thus, alongside technical and organizational risk assessments, periodic ethical risk assessments focusing on the ethical frameworks of the AIS should also be conducted. When assessing the risks associated with an AIS, it is crucial to evaluate not only technical parameters, but also ethical factors such as accountability, transparency, privacy, and freedom from unacceptable algorithmic bias. Implementing risk management practices for ethical risk assessment across all entities involved in the process is essential, despite the diverse technical activities represented by different stakeholders.

Ethical risk assessment (ERA) should be integrated into the risk management process and broaden its scope to include safety risks of an ethical nature. ERA focuses on the ethical aspects of the AIS and aids in evaluating ethical values and principles (Carrier & Narayanan, 2022). It should be conducted throughout all phases of the AIS lifecycle.

1. During *the development phase* of AIS, examples of ERA include assessing the ethical use of data, evaluating data governance, reviewing ethical issues in the data such as unfair biases, discrimination, and stigmatization, providing evidence of notifying stakeholders of potential bias issues, evaluating ethical risks associated with privacy and personal data protection, and assessing the legitimate justification for using protected or sensitive personal characteristics.

   A comprehensive risk management plan should be developed, outlining concrete processes for managing both existing and newly detected ethical risks. This plan should include control frequencies, clearly defined responsibilities, and collaboration procedures across functions. Standardized risk management practices should be implemented across all departments of an organization. In the event of failure or malfunction, documenting and sharing information is imperative to prevent future incidents.

2. In addition to the examples presented above, during *the deployment phase*, ERA involves evaluating the adequacy and appropriateness of the data used concerning transparency, accountability, and governance. It also encompasses assessing the ethical risks associated with ethical privacy, potential harms, adverse events, and unethical bias. Additionally, ERA ensures control over external system components, such as cloud services, web services, or data processing, and verifies that residual risks remain within an acceptable level of risk tolerance (Carrier & Narayanan, 2022).
3. During *the use and decommissioning phases* of an AI system, conducting assessments of the implemented risk controls becomes paramount. This involves evaluating risks associated with data transparency and sharing consent, ensuring data privacy, and averting accidental data loss. Moreover, communicating any changes in the product behavior to end-users, including updates to AI applications that may affect data exchange, storage, usage, or security, is imperative (Carrier & Narayanan, 2022). This should include an evaluation of potential disadvantages to users.

   Additionally, demonstrating consistent and predictable operational behavior of the system is crucial. Monitoring ethical issues associated with the AI system's operation, including the accessibility and behavior of the AIS to detect and mitigate any bias issues that may arise, is equally vital. Furthermore, evaluating safeguards against inadvertent loss or breaches of security and loss of control of AI data and systems, which may compromise privacy, is essential. This evaluation encompasses assessing human oversight over the functioning of mechanisms and documenting and sharing information about failure events and system malfunctions, especially safety-related data, to enable impartial investigations of AIS incidents aimed at enhancing safety standards across the entire field (Carrier & Narayanan, 2022). Moreover, assessing the disaster recovery plan and business continuity plan and managing risks and controls in the ethical decommissioning of AIS, are vital.

It is equally crucial to allocate appropriate and qualified resources, establish individual accountability for the proper operation of AIS and its outcomes, and ensure human oversight over the system operations. To effectively mitigate the ethical risks associated with AIS, it is imperative to prioritize *ethical accountability, transparency, privacy, and algorithmic bias* as fundamental criteria throughout an AIS's design, development, deployment, and operation phases. It is highly recommended to conduct thorough ethical risk assessments by qualified professionals in AI ethics who have the appropriate background and expertise.

Experience from practice and future work

The theoretical ethical requirements for risk management outlined in this paper were applied in practice during external audits. The results revealed significant gaps in implementing ethical risk management requirements. Only two out of five companies actively addressed this topic, albeit without specific processes in place. However, they demonstrated heightened attention to ethical criteria. On the contrary, the other three companies solely focused on algorithm developments, lacking knowledge of ethical requirements, and unaware of ethical standards or risk management requirements established in the AI Act or other relevant documents.

Additionally, I reached out to a total of 20 companies offering to assess their compliance with ethical requirements in AI systems development. The aim was to identify ethical risks and conduct ethical risk assessments. Out of the eight companies that responded, none had an established ethical risk assessment process or a risk management process with a focus on ethical considerations. These findings underscore a concerning lack of awareness among AI system developers regarding ethical requirements, guidelines, and standards. To address these challenges, future research should prioritize developing effective strategies to raise awareness among AI system developers and other relevant stakeholders about the necessity of complying with ethical requirements. Furthermore, investing in ongoing education and training initiatives

to enhance ethical competency and risk management capabilities is crucial for ensuring the responsible development and operation of AIS. Future research should also explore the practical implementation of regulatory risk-related requirements established in the European AI Act.

Recent independent external audits have also revealed persistent nonconformances in AI risk management despite previous internal controls and audits. Following a thematic analysis of these observations, three main areas of concerns regarding AIS risks emerged during subsequent external audits:

1. *Vulnerability to adversarial threats*: This involves the risk of system failure when faced with adversarial threats from malicious attacks. The absence of a process for documenting adversarial incidents and responses raised concerns regarding the criteria of *Robustness, Accountability, and Transparency*.
2. *Inadequate oversight and control mechanisms*: Insufficient capabilities for human oversight throughout the AIS lifecycle, coupled with a lack of established mechanisms for human intervention and mitigation of potentially harmful effects, highlighted deficiencies in the criteria of *Accountability and Transparency*.
3. *Privacy protection deficiencies*: This refers to inadequate measures to protect users' privacy, including insufficient knowledge, competency, and capability to effectively manage ethical risks and address malicious activities that may compromise privacy. These deficiencies underscored concerns related to the criteria of Privacy and Accountability.

Addressing these risks necessitates comprehensive measures to enhance AIS robustness, strengthen accountability and transparency mechanisms, and reinforce privacy protection procedures. Further research should also explore effective control methods, such as independent external ethical audits of high-risk AI systems.

Limitations of this research include the limited sample size employed and the challenge of enforcing compliance with ethical standards and risk management, particularly due to the pending enactment of the AI Act. As a result, enlarging the sample size at this stage would lack justification and the research will need to be revisited once the AI Act is enforced.

Root cause analysis and corrective and prevention actions

Root cause analyses are a fundamental component of the risk management process, typically conducted after identifying deficiencies during audits. A root cause signifies a foundational deficiency flaw leading to nonconformance, necessitating correction to prevent the recurrence of similar issues. A root cause analysis is applicable across all phases of the process and should be promptly initiated upon identifying any risk. It involves evaluating various areas:

- *Technical parameters, data, and AI systems:* Assessing factors such as the quality and accuracy of information and data, the integrity of AI models, and the robustness of algorithms.
- *Organizational processes:* Identifying communication gaps, deficiencies in training, lapses in delegation of responsibilities, and inadequacies in human oversight.
- *Ethical requirements:* Ensuring adherence to ethical principles encompassing respect for human rights and dignity, privacy and personal data protection, fairness and non-discrimination, individual, social, and environmental well-being, as well as transparency and accountability. Ethical requirements also outline risk mitigation strategies to safeguard core values.

It is imperative to develop an intervention plan to initiate corrective interventions within the requisite timeframe if the entire AI system or any of its components fail to comply with the organization's standard operating procedures or ethical standards. Furthermore, establishing procedures for suspending activities and possessing appropriate competencies to halt operations when necessary is paramount.

Upon identifying the root cause of an issue during the analysis, implementing appropriate Corrective and Preventive Actions (CAPAs) becomes pivotal. Corrective actions are aimed at eliminating the root cause(s) of *existing* deviations, nonconformities, defects, or other undesirable situations to prevent their recurrence. Conversely, preventive actions are designed to address the root cause(s) of *potential* deviations, nonconformities, or defects to avert their occurrence in the future. Subsequently, conducting a post-implementation effectiveness assessment of CAPAs is necessary to ascertain their efficacy in preventing recurrence and evaluating the likelihood of the same nonconformity resurfacing.

Conclusion

Risk management plays a key role in conducting ethical assessments of AI, facilitating the identification, analysis, and mitigation of potential negative ethical impacts associated with AI applications. By adhering to proper risk management practices, organizations establish a systematic approach to AI risk governance, ensuring consistent and efficient achievement of desired outcomes. Ethical considerations integrated into risk management are fundamental across the entire lifecycle of AI. This integration protects privacy, fosters trust, enhances cybersecurity, mitigates unfair biases, and promotes transparency and accountability in AI systems. While risk management practices may vary depending on context and use, they all share the same objective of embedding values-based principles into AI systems through procedural and technical attributes. Risk management is indispensable for ensuring the safety of AI systems which is crucial for fostering trust in emerging technologies.

Establishing mechanisms for transparency and accountability is essential for ensuring the trustworthiness of AI systems and their outcomes. Stakeholders involved in the development and utilization of AI systems should bear verifiable accountability for their proper functioning. Auditability emerges as a crucial element in ensuring the verifiable security of stakeholders impacted by AI systems, enabling comprehensive assessments across various dimensions such as algorithms (including opaque algorithms), data governance (including biased data sets), design processes (including adherence to ethical principles and correlated values), system performance and functionality, and respective accountabilities (including effective risk governance). AI audits should verify compliance with key requirements and applicable procedural standards that AI systems should meet to be deemed trustworthy and consistent with ethical principles. Any nonconformances and compliance concerns observed during AI audits should be adequately addressed to minimize potential safety, regulatory, or business risks. Schuett (2023) suggested that regulations should mandate the evaluation of risk management effectiveness through internal audits. However, discussions with relevant stakeholders and both internal and external auditors reveal that external audits provide more objective and valuable insights into improving the effectiveness of controls, operations, risk management, and governance processes. The author of this paper audited all phases of the risk management process outlined herein. The analysis of audit results indeed demonstrates that external audits are more effective in assessing risk management procedures than internal audits. During external audits, risk-based observations were detected even though internal audits had been previously conducted. External auditors can identify risk-based issues that internal auditors might overlook due to professional blindness or automated acceptance of existing procedures, potentially leading to ineffective solutions and regulatory or market repercussions. Therefore, the author recommends the inclusion of a requirement for independent external audits to evaluate the effectiveness of organizations' risk management systems. This constitutes a crucial contribution to the ongoing experts' discussion on consistently, efficiently, and satisfactorily achieving desired outcomes while exceeding stakeholders' expectations in diverse circumstances. Ethical risk assessment activities necessitate a diverse range of specialist skills and expertise, requiring experienced professionals with a background in AI ethics.

The AI Act establishes a penalty system for companies found to be non-compliant with its regulations. Fines for violations of the AI Act were set as a percentage of the offending company's global annual turnover in the previous financial year or a predetermined amount, whichever is higher. Specifically, fines amount to €35 million or 7% for violations involving banned AI applications, €15 million or 3% for violations of other obligations outlined in the AI Act, and €7,5 million or 1% for supplying incorrect information (AI Act, 2023b).

The overview of AI risk management presented in this paper can serve as a foundation for further explorations. This is particularly important in the development of AI risk management software, which could become a standard tool for simplifying risk management process while addressing ethical risks in AI. Several new methodologies for measuring cyber risks, with potential utility in guiding the implementation of preventive actions, have already been studied and proposed by several authors (Facchinetti, Giudici & Osmetti, 2020; Giudici & Raffinetti, 2022; Zängerle & Schiereck, 2023). Giudici et al. (2024) introduced an AI risk management framework focused on the assessment of AI risks. This framework, built upon the four fundamental statistical principles of SAFEty - Sustainability, Accuracy, Fairness, and Explainability, presents a methodology for evaluating the safety and trustworthiness of AI applications. For each principle, the authors proposed a set of interconnected statistical metrics named Key AI Risk Indicators (KAIRI). These metrics not only facilitate the evaluation of AI applications' safety and trustworthiness but also enable continuous monitoring to uphold their integrity.

Acknowledgement
This work was partially supported by the Ministry of the Interior of the Czech Republic under the project identification code VI04000107 and by the Cooperatio Program, research area Philosophy, Charles University, Faculty of Humanities.

References

AI ACT (2023a): *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonized rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. [online] [Retrieved December 18, 2023] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

AI ACT (2023b): *Artificial Intelligence Act: Council and Parliament strike a deal on the first rules for AI in the world*. [online] [Retrieved December 18, 2023] Available at: https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/aFrtificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/

BAQUERO, J. A., BUKHAARDT, R. et al. (2020): *Derisking AI by design: How to build risk management into AI development*. [online] [Retrieved February 28, 2024] Available at: https://www.mckinsey.com/capabilities/quantumblack/our-insights/derisking-ai-by-design-how-to-build-risk-management-into-ai-development

CARRIER, R. & NARAYANAN, S. (2022): FH – AI risk management process. In: *ForHumanity*. [online] [Retrieved December 18, 2023] Available at: https://docs.google.com/document/d/1tsFWSjQlE5kcyfLqlbTBPnGD8opsYNuSciROMi0yFOM/edit

ELING, M., McSHANE, M. & NGUYEN, T. (2021): Cyber risk management: History and future research directions. In: *Risk Manag Insur Rev.*, 24, pp. 93–125. DOI: 10.1111/rmir.12169.

EUROPEAN PARLIAMENT (2020): *Artificial framework of ethical aspects of artificial intelligence, robotics and related technologies*. [online] [Retrieved December 18, 2023] Available at: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.pdf

FACCHINETTI, S., GIUDICI, P. & OSMETTI, S. A. (2020): Cyber risk measurement with ordinal data. In: *Stat Methods Appl*., 29, pp. 173–185. DOI: 10.1007/s10260-019-00470-0.
FRASER, H. L. & BELLO Y VILLARINO, J. M. (2021): Where residual risks reside: A comparative approach to Art 9(4) of the European Union's proposed AI regulation. In: *SSRN*. DOI: 10.2139/ssrn.3960461.
GIUDICI, P., CENTURELLI, M. & TURCHETTA, S. (2024): Artificial intelligence risk measurement. In: *Expert Systems with Applications*, 235, 121220. DOI: 10.1016/j.eswa.2023.121220.
GIUDICI, P. & RAFFINETTI, E. (2022): Explainable AI methods in cyber risk management. In: *Qual Reliab Eng Int*., 38, pp. 1318–1326. DOI: 10.1002/qre.2939.
HAN, S., KELLY, E., NIKOU, S. & SVEE, E. O. (2022): Aligning artificial intelligence with human values: reflections from a phenomenological perspective. In: *AI & Soc*., 37, pp. 1383–1395. DOI: 10.1007/s00146-021-01247-4.
CHAMBERLAIN, J. (2023): The risk-based approach of the European Union's proposed artificial intelligence regulation: Some comments from a tort law perspective. In: *European Journal of Risk Regulation*, 14(1), pp. 1–13. DOI: 10.1017/err.2022.38.
IEEE (2001): *Standard for software life cycle processes – risk management*. [online] [Retrieved December 18, 2023] Available at: https://standards.ieee.org/ieee/1540/2280/
IEEE (2021): IEEE standard model process for addressing ethical concerns during system design. In: *IEEE Std 7000-2021*. DOI: 10.1109/IEEESTD.2021.9536679.
ISO 31000 (2018): *ISO 31000 risk management – guidelines*. Second edition. [online] [Retrieved December 18, 2023] Available at: https://shahrdevelopment.ir/wp-content/uploads/2020/03/ISO-31000.pdf
KPMG (2021): *Artificial Intelligence in Risk Management*. [online] [Retrieved February 28, 2024] Available at: https://kpmg.com/ae/en/home/insights/2021/09/artificial-intelligence-in-risk-management.html
KUMAR, R., SHANKAR, S. & NAGLE, F. (2020): The Case for AI Insurance. In: *Harvard Business Review Digital Articles.* [online] [Retrieved February 28, 2024] Available at: https://hbr.org/2020/04/the-case-for-ai-insurance
MACRAE, C. (2022): Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk. In: *Risk Anal*., 42(9), pp. 1999–2025. DOI: 10.1111/risa.13850.
MADIEGA, T. (2023): Artificial intelligence act. In: *EU Legislation in Progress*. [online] [Retrieved December 18, 2023] Available at: https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf
NIST (2023): *The artificial intelligence risk management framework*. DOI: 10.6028/NIST.AI.100-1.
OECD (2023): Advancing accountability in AI: Governing and managing risks throughout the lifecycle for trustworthy AI. In: *OECD Digital Economy Papers*, 349. DOI: 10.1787/2448f04b-en.
PATÉ-CORNELL, M. E., KUYPERS, M., SMITH, M. & KELLER, P. (2018): Cyber risk management for critical infrastructure: A risk analysis model and three case studies. In: *Risk Analysis*, 38(2), pp. 226–241. DOI: 10.1111/risa.12844.
POEL, I. (2020): Embedding values in artificial intelligence (AI) systems. In: *Minds & Machines*, 30, pp. 385–409. DOI: 10.1007/s11023-020-09537-4.
SCHUETT, J. (2023): Risk management in the Artificial Intelligence Act. In: *European Journal of Risk Regulation*, pp. 1–19. DOI: org/10.1017/err.2023.1.
UNESCO (2022): *Recommendation on the ethics of artificial intelligence*. [online] [Retrieved December 18, 2023] Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137

ZÄNGERLE, D. & SCHIERECK, D. (2023): Modelling and predicting enterprise-level cyber risks in the context of sparse data availability. In: *Geneva Pap Risk Insur Issues Pract.*, 48, pp. 434–462. DOI: 10.1057/s41288-022-00282-6.