Funded by the European Union

secureu
DIGITAL EDUCATION TOOLS FOR SECURITY RISK MANAGEMENT

AI SECURITY CHALLENGE AND RISK ASSESSMENT USING ISO 31000: THE IOTSI GUIDANCE

Rita Lankauskienė/ Kazimieras Simonavicius University, Lithuania /2024

ABSTRACT

The expansion of generative Artificial Intelligence (AI) technologies across various sectors presents substantial advantages while simultaneously posing intricate security challenges. Managing these risks effectively is essential for maintaining the integrity, reliability, and safety of AI systems. The ISO 31000 standard offers a systematic methodology for risk management that may be tailored to the particular requirements of AI security. This article outlines the emerging cyber security issues with AI challenges at the forefront, which occur due to the unpredictably rapid expansion of the Internet of Things (IoT) industry, accelerated by digital transformation. Based on the Internet of Things Security Institute's (IoTSI) best practice, a step-by-step explanation is given of how to empower the ISO 31000 standard in an AI security risk assessment, encompassing comprehensive methods, technical analysis, and illustrative examples.

Link to ISO 31000

This article explores the application of the ISO 31000 standard systematic methodology and process framework for generative Artificial Intelligence (AI) Security Risk Assessment in organizations.

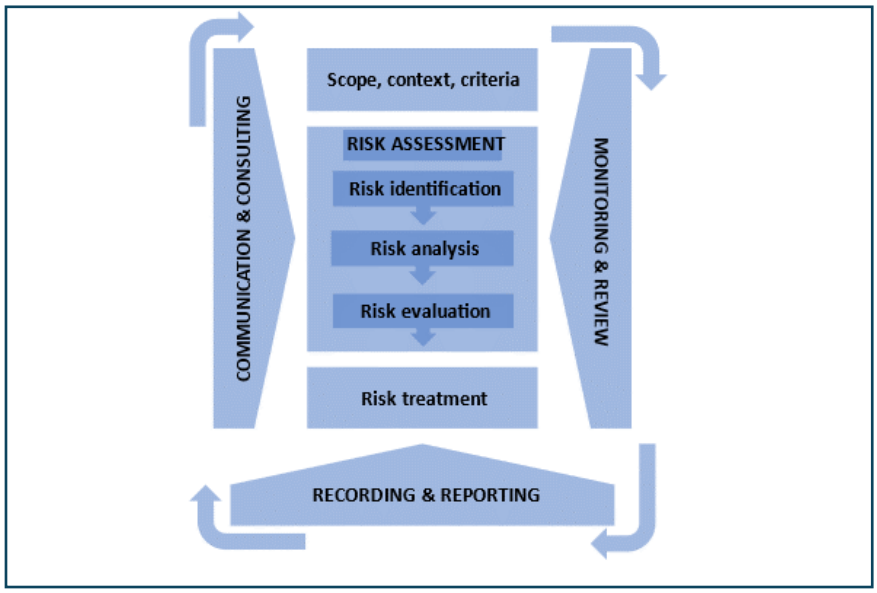


Figure 1. Risk management process according to ISO 31000:2018

1. Introduction

Emerging technologies with generative Artificial Intelligence (AI) at the forefront have widely become recognized as significant catalysts for digital transformation and innovation worldwide (Thorat et al., 2024). Plenty of evidence is collected proving the tremendous benefits, gathered from the numerous applications of various AI tools in different spheres of human activity and technological developments (e.g., Ghobakhloo et al., 2024; Kanbach et al., 2024; Sedkaoui & Benaichouba, 2024, etc.).

The swift development of AI technology and the widespread use of creative AI solutions have led to the rapid emergence of new risk types, which further increases the unpredictability of the already complex procedures involved in AI creation and implementation (Golpayegani et al., 2022). The increasing number of incidents (AI Incident Database, 2024; AIAAIC, 2024) brought on by the (mis)use of AI increasingly bothers a broad range of stakeholders from the general public, organizations, and governments at different – national, international, and global levels (Herani & Angela, 2024).

Numerous initiatives are being undertaken internationally, as well as globally, to limit the negative effects of AI, ranging from standards for risk management and regulatory frameworks to guidelines encouraging reliable development and use. Among the most relevant recently issued regulatory frameworks in Europe is the European Commission's Artificial Intelligence Act (AI Act, 2019). The AI Act (2019) is the world's first all-encompassing legal framework around AI. The regulations seek to promote trustworthy AI in Europe and beyond, aiming to guarantee that AI systems adhere to fundamental rights, safety, and ethical standards, and addressing the risks associated with extremely potent and significant AI models.

The core premise behind the AI Act (2019) is that it guarantees Europeans confidence in the potential of AI. Certain AI systems present risks that should be managed to prevent unfavorable consequences, even if the majority of AI systems are low- to non-risk and can help solve many social issues. For instance, it is frequently impossible to determine the rationale behind an AI system's decision, forecast, or action. Therefore, determining whether someone has been unfairly disadvantaged - for example, in an employment decision or during an application for a public benefit program - may become challenging.

Even if current laws offer some protection, they are not enough to handle the unique difficulties that AI systems can present. The AI Act is well-known because it regulates the development and application of AI in systems in a risk-oriented manner. By implementing a risk management system for risk detection, analysis, evaluation, and treatment, risk management practices seek to handle core uncertainties. In this case, the uncertainties of AI systems and their dangers are to be handled following ISO risk management standards. There is collected evidence on how the safe and reliable development and use of AI systems depend on the proper implementation of ISO 31000 in any entity‘s risk management activities.

The ISO 31000 series of standards offers actions, guidelines, and principles to help organizations manage risk. The primary standard that offers general guidelines, a framework, and procedures for handling risks that organizations encounter during their lifecycle is ISO 31000:2018 Risk Management - Guidelines. Another relevant member of this family is ISO 31073:2022 Risk Management – Vocabulary (2022), which enables a common understanding across various business units and organizations, ISO 31073:2022 offers a list of general terms in risk management together with their meanings.

This article presents guidance on best practise, developed by the Internet of Things Security Institute (IoTSI), on how to implement a step-by-step AI security risk assessment procedure in an organization, using the ISO31000 framework.

2. Case

Digital transformation results appear in the unpredictably rapid expansion of the Internet of Things (IoT) industry. More and more devices becoming connected and providing new features and increased convenience in both the personal and professional spheres. According to the IoT Analytics' data (2024), by the end of 2023, 16.6 billion IoT devices were connected, which was a 15% increase over 2022. And by the end of 2024, IoT Analytics (2024) projects that this will have increased by 13% to 18.8 billion, and will keep increasing further (see figure 2).

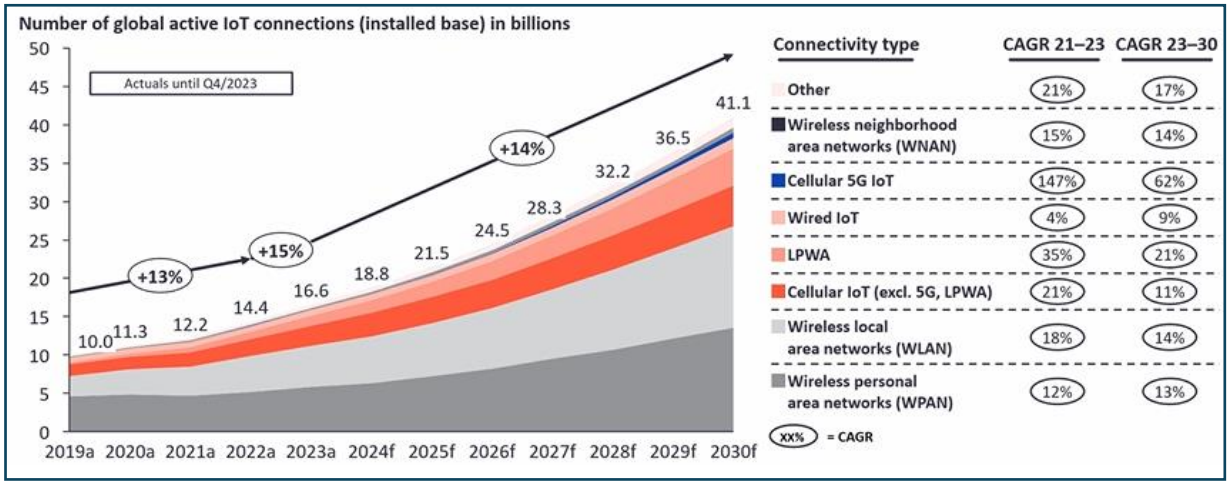
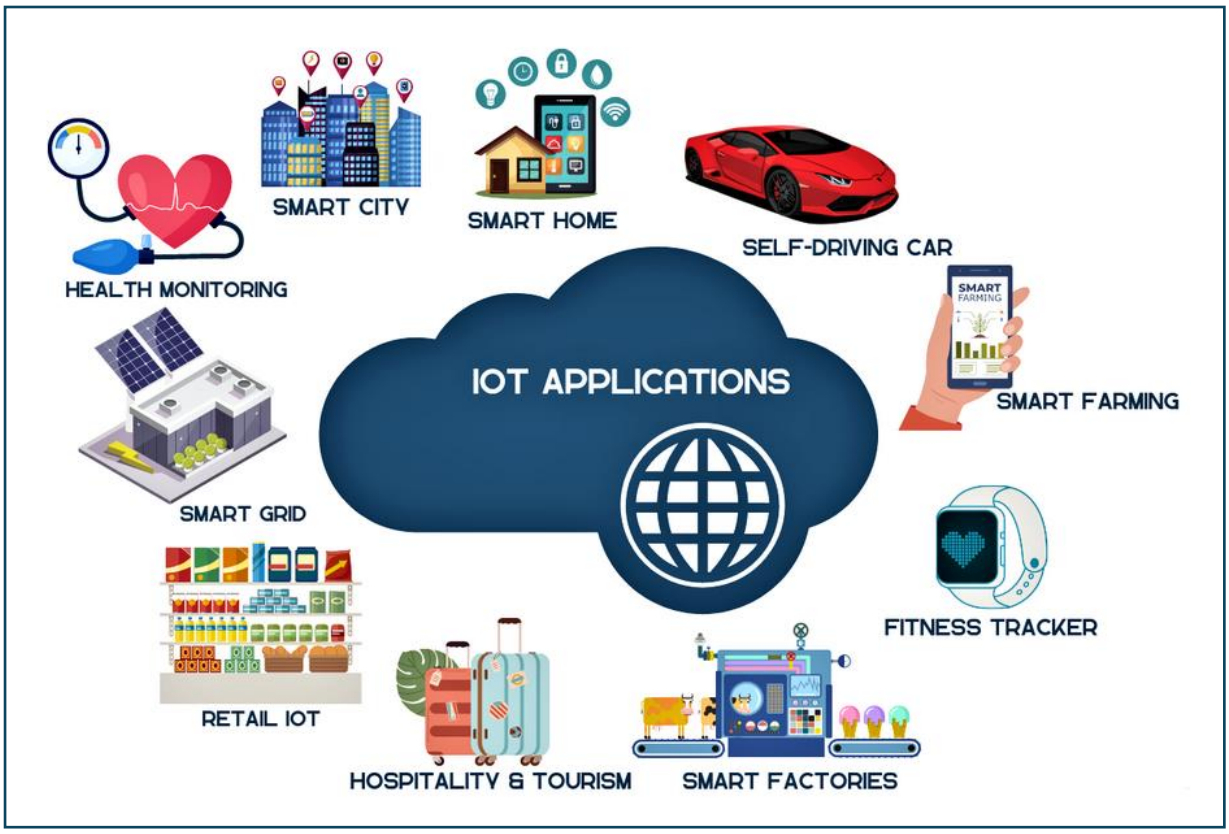


Figure 2. IoT Analytics, State of IoT Summer 2024. Available at: https://iot-analytics.com/product/state-of-iot-summer-2024/

Alongside the growing number of IoT devices and systems, their complexity and sophistication are also rising, offering numerous advantages to a wide range of sectors, such as manufacturing, healthcare, smart cities, home automation, and many more (see picture 1).



Picture 1. Internet of Things (IoT) applications' areas - AI security challenge. Source: Techjury.net, 2024.Available at: https://techjury.net/blog/internet-of-things-statistics/

At the same time, IoT application areas and the widespread adoption of AI technology in many sectors yield substantial advantages while simultaneously presenting intricate security issues. According to the best practice in the field, the above-outlined security issues are successfully manageable following the principles embedded in the ISO 31000 standards' family.

Due to the increasing complexity of the issue, a particular academic and cyber industry think tank has been established - the Internet of Things Security Institute (IoTSI, 2024). IoTSI is focused on delivering security frameworks, instructional resources, and cybersecurity courses to facilitate the best practices in the management of security in Smart Tech, IoT (Internet of Things), and IIoT (Intelligent Internet of Things) ecosystems.

IoTSI suggests using the ISO 31000 standard's systematic framework for risk management, which may be well tailored to the particular requirements of AI security. It is further described in detail, how to apply the ISO 31000-based methodology for an AI security risk assessment, encompassing comprehensive methods, technical analysis, and illustrative examples.

3. Best practices

ISO 31000 is an international standard that establishes risk management guidelines to guarantee that risks are managed consistently and systematically at all organizational levels (ISO31000:2018). It comprises three primary elements: 1) principles, 2) framework, and 3) process. The framework provides the structural elements for implementation, the principles ensure that risk management is a component of decision-making and adds value, and the process entails steps for identifying, assessing, and mitigating risks.

According to IoTSI guidance, an AI security risk assessment, using ISO 31000, is a step-by-step procedure, which consists of several complementary steps (see figure 3).

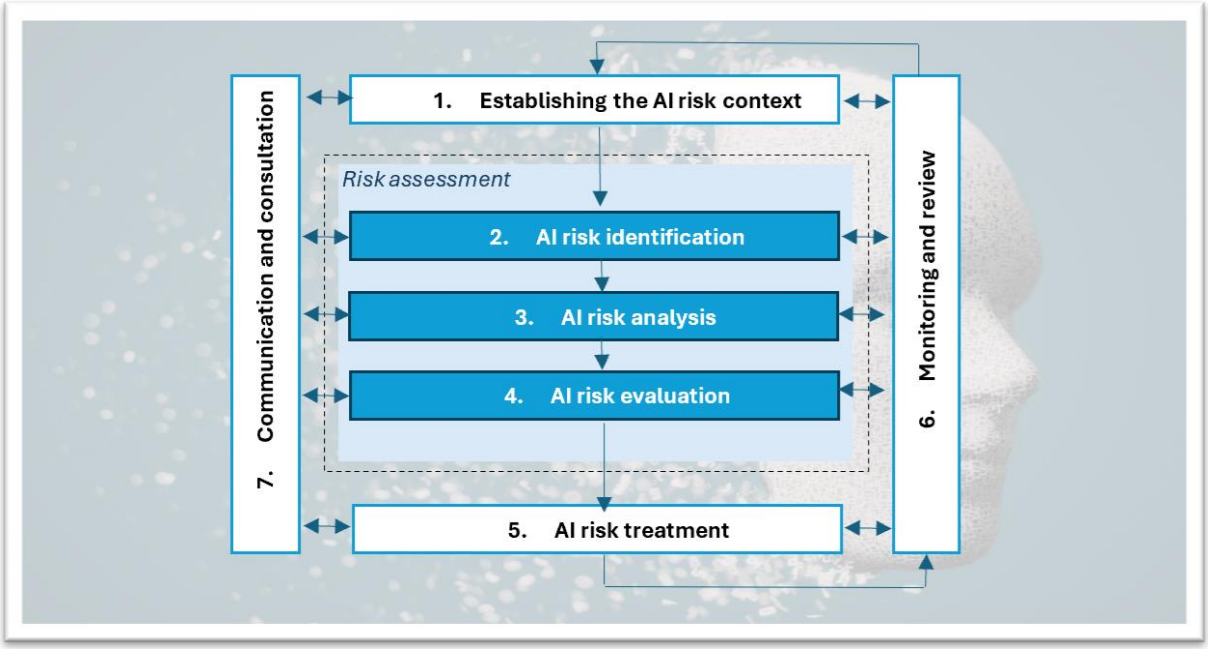


Figure 3. AI security risk assessment, using ISO 31000. Source: Adapted from IoTSI guidance (2024).

3.1. Establishing the context

The foundational step in the ISO 31000 process is the establishment of context, which provides the necessary background to comprehend the environment in which the AI system operates. It is necessary to consider the environment from the two core contexts: external and internal.

While establishing the internal context, it is necessary to focus on the 3 core things:

1) Regulatory landscape - find the rules and laws that relate to data handling, privacy, and security, such as GDPR, CCPA, and HIPAA. As an example, GDPR rules must be followed by any AI system that handles personal data in the EU.

2) Market and technological trends - learn about the changes in the market and in technology that might affect the security of the AI system. For instance, if adversarial attack methods get better, defenses may need to be updated.

3) Threat landscape - look at the current danger landscape. This should include known cyber threats that target AI systems, like data poisoning or model inversion attacks.

The external context refers to the following core themes:

1) Organizational structure - explain who is responsible for what when it comes to managing and protecting AI systems. Find the important people, like IT, legal, safety, and business units.

2) Risk management policies - check the current risk management policies to see if they match the risks that come with AI and the ISO 31000 rules.

3) Risk Appetite and Tolerance - name the company's risk appetite and tolerance levels, especially when it comes to system security, data breaches, and theft of intellectual property.

Once the environmental context is established, it is necessary to define the risk criteria. You should come up with specific ways to assess risks. For example, you may first define the impact metrics, such as financial loss, reputational damage, operational disruption, and legal consequences. Then consider the likelihood metrics, including frequency of threat occurrence, vulnerability exploitability, and control effectiveness. And finally, it is necessary to establish risk categorization i.e., low, medium, and high, based on the above – their impact and likelihood.

3.2. AI Risk Assessment Procedure

The next step is AI Risk Assessment (see picture 2), which is composed of several components. ISO 31000 lists three main components for assessing risk: risk identification, risk analysis, and risk evaluation. The goal of this part is to fully understand all the possible threats and weak spots in the AI system.

3.2.1. AI risk identification

In the risk identification phase, it is necessary to consider all potential sources of risk, including data risks, model risks, and operational risks.

Data risks consist of data breaches, data poisoning, and data integrity. Data breaches are related to unauthorized access to private or secret data, which could result in fines and a loss of trust. Data poisoning is hacking training data to change how an AI model acts in a bad way. For example, changing how accurate a spam blocker is by adding false data to its training set. Data integrity is a risk that comes from the need for correctness and thoroughness of the data, which can affect how well and reliably the AI model works.

Model risks entail adversarial attacks, model stealing, and model bias. The adversarial attacks cause changes to input data that are meant to trick the AI model, like changing pictures to force an image recognition system to make wrong decisions. Model stealing occurs when hackers can get into the structure or settings of an AI model without permission, which can lead to theft of intellectual property or the creation of competing models. Model bias becomes relevant since AI models can have unintended biases that can lead to unfair results. This is especially important in sensitive areas like loans or hiring.

Operational risks overwhelm system failures, security configuration, and third-party risks. System failures come with hardware or software problems that can stop an AI system from working. Security configuration causes problems with the AI system's security, like not enough access controls or gaps that haven't been fixed, which can let attackers in. Finally, third-party risks come from sellers or third-party services, like cloud providers, which can make data less safe and systems less available.

3.2.2. AI risk analysis

The essence of risk analysis is to look carefully at each risk you've found in the previous stage, to learn what it is, how it might affect you, and how likely it is to happen, i.e. perform an impact assessment and likelihood assessment. It is advised by best IoTSI practice to find out about risks by using both numeric and qualitative methods.

During the impact assessment, first consider the financial impact by figuring out how much data breaches could cost in terms of fines, court fees, and fixing problems. Then consider operational impacts, i.e., think about how system downtime, loss of usefulness, and changes to business processes might affect things. And finally analyze reputational impact, considering how this will affect the company's reputation, customer trust, and position in the market over the long run.

In the likelihood assessment, first, consider the historical data, by looking at past events and weaknesses to figure out how likely it is that they will happen again. Then perform the vulnerability analysis by assessing the AI system's exposure to identified threats, considering factors like the system's complexity and the robustness of existing security measures. And finally, analyze threat actor capability, i.e., evaluate the capability and motivation of potential attackers, such as hackers, malicious insiders, or competitors.

For example, to assess the danger of adversarial assaults on sensor data in an AI-based autonomous car system, take into account the impact on passenger safety and system reliability.

3.2.3. AI risk evaluation

The following risk evaluation procedure is implemented by ranking the risks and comparing the ones that have been studied to the set standards for risks. This helps decide where to put resources and what risks are the most important.

IoTSI best practices advice using the risk matrix for sorting the risks into groups based on how likely they are to happen and how bad they could be. This visual tool helps you put risks in order of importance and make smart choices about how to treat risks. The decision-making will be more effective by getting everyone involved in figuring out what amounts of risk are acceptable and how to prioritize risks. When making decisions, people may have to weigh the costs and benefits of reducing danger.

3.3. Risk Treatment

Risk treatment is the process of choosing and putting into action ways to change dangers. ISO 31000 says that this can mean avoiding, transferring, reducing, or taking risks.

To manage the process, it is necessary to develop risk treatment plans. A thorough risk treatment plan should be developed for each significant risk. From the best practices, each plan should define the objective, actions, and resources and allocate responsibilities. The objective must define clearly what a particular treatment is supposed to do, like lowering the risk of data leaks or weakening the effects of hostile attacks. The actions must be specified to tell the people what they need to do, like putting in place multi-factor login, encrypting data, or doing regular security checks. It is necessary to equip the risk treatment measures with reasonable resources, staff, and technology they need to be put into action. And finally, every plan must clearly assign responsibilities for executing the treatment plan, ensuring accountability and oversight.

Risk treatment plan example: To mitigate the risk of model bias, a treatment plan may encompass diversifying training data, applying fairness metrics, and performing regular audits to identify and rectify biases.

The next step of risk treatment plan implementation entails putting the risk treatment steps into action and making sure they are part of how the organization works and how it is managed. Technical controls usually are performed by using high-tech security tools like encryption, firewalls, intrusion detection systems, and programmes that look for strange behaviour. Procedural controls are ensured by setting up or improving processes for managing data, controlling who can see it, and responding to incidents. For example, setting up a way to check for and update security patches daily. And organizational controls are devoted to creating a mindset of security awareness through training programmes, campaigns, and making sure that the IT and business units work together.

Risk treatment plan implementation example: establishing a zero-trust security framework for an AI-driven cloud service, incorporating stringent access limits, ongoing surveillance, and thorough audit trails.

3.4. Monitoring and review

Regular reviews and ongoing monitoring are necessary to make sure that risk management methods keep working and adapt to new risks.

First, best practices highlight the necessity of establishing a way to keep an eye on the AI system and its surroundings all the time so that you can quickly spot and deal with new risks. Security Information and Event Management (SIEM) solutions are used to gather and study data about security, sending alerts in real-time for possible security incidents. Another essential part is the model monitoring to continuously monitor the performance of the AI model all the time, aiming to find strange things, like output patterns that don't make sense, which could be a sign of an attack or model change. Finally, incident management is used for setting incident response plans with roles, communication channels, and recovery steps to deal with security incidents quickly.

Continuous monitoring example: using AI-based monitoring tools to find strange trends of data access in a financial AI system could mean that there are threats from inside the company or outside the company.

To ensure a well-running monitoring system, periodic reviews are necessary. The best practice advice is to review the treatment plans and risk management process often to make sure they are still useful and effective. This is performed by internal audits, scenario analyses, and stakeholder engagement. It is advised to do internal audits to see if risk management methods are adequate and working well. This includes checking to see if the rules and policies of the company are being followed. Scenario analysis is useful when testing how well the AI system can handle made-up threats like a coordinated cyberattack or a big data breach, you should do scenario analysis. Stakeholder engagement is helpful for reviews: talking to stakeholders helps reconsider the results of risk management, make changes to risk factors, and improve plans for risk treatment.

Periodic review example: evaluating and revising the risk evaluation of an AI-driven healthcare diagnostic instrument following recent regulatory directives regarding patient data privacy.

3.5. Communication and consultation

Effective communication and consultation are essential components of the ISO 31000 process, guaranteeing transparency and stakeholder engagement. The best practice in the field takes into account both internal and external communication.

Internal communication deals with risk management actions and results, that should be shared regularly with all internal stakeholders that need to know, such as the board of directors, management, and operational staff. Reporting is necessary for giving thorough reports on incidents, risk assessments, and efforts to lower risks. It advised to include important metrics like the level of risk, how well the controls are working, and the state of compliance. And finally, internal communication is concerned with training and awareness. It is necessary to teach the staff regularly about best practices for security, new threats, and how they can help control risk.

Internal communication example: holding meetings for engineers and data scientists to talk about safe ways to build AI and how important it is to think about data quality and ethics.

External communication is used to tell outside groups, like customers, partners, and regulatory bodies, about the efforts to handle risks and follow the rules. It is important to consider the transparency principle, i.e., be clear about the AI system's security measures, especially when it comes to privacy and data protection. Another important part is reporting incidents. It is necessary to establish rules for telling regulators and people touched by security incidents about them, making sure that the information is given correctly and on time.

External communication example: after a security breach in a consumer app driven by AI, a public statement should be made explaining the steps that were taken to protect user data and stop future breaches.

3.6. Recording and reporting

For accountability, openness, and constant growth, it's important to keep detailed records and reports. Documentation is the first important component of this. Best practices advise to write down everything that you do during the risk management process, such as finding risks, analyzing them, making treatment plans, and keeping an eye on things. To create records management, it is necessary to set up a method for managing records to store and organize paperwork so that it is easy to find and that you are following the law. Audit trails are applied to maintain full records of all risk management activities, such as choices made, actions taken, and results reached so that they can be easily checked.

Documentation example: develop comprehensive documentation of a risk assessment procedure for an AI system, encompassing data sources, risk criteria, analytical methods, and mitigation options.

Reporting is based on preparing regular reports: periodic reports and compliance reports. They are used to inform stakeholders about the status of AI security risk management. It is advised to produce regular periodic reports that summarize the most important risks, how they are being managed, and security events, including measurements and trends to get a full picture of the risk exposure. Compliance reports help ensure that you follow all the rules and laws that apply by writing reports that show you did so, like data security rules.

Reporting example: generating a yearly risk management report for an AI-driven financial system, emphasizing significant risks, the efficacy of controls, and opportunities for enhancement.

Summing up, the best practice, proposed by IoTSI for conducting AI security risk assessment using ISO 31000, is worth applying for several reasons. The ISO 31000 framework offers a comprehensive and methodical strategy for addressing the distinct security threats linked to AI systems. Organizations can reduce potential dangers, guarantee regulatory compliance, and protect their AI assets by adhering to a disciplined approach that includes establishing context, conducting risk assessments, executing treatment plans, and regularly monitoring and reviewing. Efficient communication and comprehensive documentation enhance transparency, accountability, and ongoing enhancement in AI security risk management. As AI technologies and their associated risks develop, it will be imperative for enterprises to implement a comprehensive risk management framework such as ISO 31000 to effectively navigate the intricate environment of AI security.

References

AI Act (2019). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689.

AI Incident Database (2024). Available at: https://incidentdatabase.ai/.

AIAAIC - AI, Algorithmic and Automation Incident and Controversy Repository (2024). Available at: https://www.aiaaic.org/home.

Conducting an AI security risk assessment using ISO 31000 (2024). IoTSI. Available at: https://iotsecurityinstitute.com/iotsec/index.php/iot-security-institute-blog/155-conducting-an-ai-security-risk-assessment-using-iso-31000.

Ghobakhloo, M., Fathi, M., Iranmanesh, M., Vilkas, M., Grybauskas, A., & Amran, A. (2024). Generative artificial intelligence in manufacturing: opportunities for actualizing Industry 5.0 sustainability goals. *Journal of Manufacturing Technology Management*, *35*(9), 94-121.

Golpayegani, D., Pandit, H. J., & Lewis, D. (2022). Airo: An ontology for representing AI risks based on the proposed EU AI act and ISO Risk management standards. In *Towards a Knowledge-Aware AI* (pp. 51-65). IOS Press.

Herani, R., & Angela, J. (2024). Navigating ChatGPT: catalyst or challenge for Indonesian youth in digital entrepreneurship?. *Journal of Entrepreneurship in Emerging Economies*. Vol. ahead-of-print No. ahead-of-print. Available at: https://doi.org/10.1108/JEEE-05-2024-0181.

IoT Analytics. State of IoT, Summer 2024. Market Report. Available at: https://iot-analytics.com/product/state-of-iot-summer-2024/.

IoTSI (Internet of Things Security Institute) (2024). Available at: https://iotsecurityinstitute.com/iotsec/index.php/about.

ISO 31000:2018. Risk management – Guidelines (2018). Available at: https://www.iso.org/obp/ui/en/#iso:std:iso:31000:ed-2:v1:en .

ISO 31073:2022 Risk Management – Vocabulary (2022). Available at: https://www.iso.org/obp/ui/en/.

Kanbach, D. K., Heiduk, L., Blueher, G., Schreiter, M., & Lahmann, A. (2024). The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, *18*(4), 1189-1220.

Petrov Ch. (2024). 26 Insightful Internet of Things Statistics 2024. Techjury.net. Available at: https://techjury.net/blog/internet-of-things-statistics/.

Sedkaoui, S., & Benaichouba, R. (2024). Generative AI as a transformative force for innovation: a review of opportunities, applications and challenges. *European Journal of Innovation Management*, Vol. ahead-of-print No. ahead-of-print. Available at: https://doi.org/10.1108/EJIM-02-2024-0129.

Thorat, S. R., Tingare, B. A., Deshmukh, S. R., Dabhade, V. D., William, P., Rakshe, D. S., & Verma, A. (2024). Analysis Of Generative Ai's Impact On Industry 4.0 And Digital Transformation. *Library Progress International*, *44*(3), 13379-13390.