2022 International Conference on Advanced Enterprise Information System (AEIS) | 979-8-3503-2437-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/AEIS59450.2022.00017

AIRMan: An Artificial Intelligence (AI) Risk Management System

Simon Tjoa, Peter Kieseberg Marlies Temper
Institute of IT Security Research
St. Pölten University of Applies Sciences
St. Pölten, Austria
e-mail: <name>.<surname>@fhstp.ac.at

Jakob Zanol
Centre for Computers and Law
University of Vienna
Vienna, Austria
e-mail: jakob.zanol@univie.ac.at

Markus Wagner
Institute of Creative Media Technology
St. Pölten University of Applies Sciences
St. Pölten, Austria
e-mail: markus.wagner@fhstp.ac.at

Andreas Holzinger
University of Natural Resources and Life Sciences
Institute of Forest Engineering (FT)
Vienna, Austria
e-mail: andreas.holzinger@boku.ac.at

***Abstract*— Artificial intelligence (AI) has emerged as one of the most formative technologies of the century and further gains importance to solve the big societal challenges (e.g. achievement of the sustainable development goals) or as a means to stay competitive in today's global markets. The role as a key enabler in many areas of our daily life leads to a growing dependence, which has to be managed accordingly to mitigate negative economic, societal or privacy impacts. Therefore, the European Union is working on an AI Act, which defines concrete governance, risk and compliance (GRC) requirements. One of the key demands of this regulation is the operation of a risk management system for High-Risk AI systems. In this paper, we therefore present a detailed analysis of relevant literature in this domain and introduce our proposed approach for an AI Risk Management System (AIRMan).**

Keywords- Artificial Intelligence; AI; Security; Risk Management; AI Act; High Risk AI; AI Risk Management System, Governance; Compliance

I. INTRODUCTION

Artificial intelligence (AI) is indisputably one of the most important and formative technologies today. Across all industries and countries AI is applied to enable new products & services (e.g. self-driving cars), to make organizational processes more efficient (e.g. predictive maintenance) or to support daily activities (e.g. personal assistants).

However, the widespread use of AI led to the call for new rules and regulations in order to ensure that the technology is developed and operated in a trustworthy and ethical manner [1], especially with respect to the problem of missing explainability of such systems that further complicates standard strategies for ensuring a certain level of security [2]. Therefore, the European Union started a debate on how to ensure that technological advancement in this area can be regulated for critical areas and at the same time not impede the development of this important technology. This led to the proposal of the AI Act [3].

The introduction of a working risk management for high-risk AI systems is one of the major pillars in the current AI Act proposal. Also, with respect to other obligations imposed on operators of such systems by the act, the ability to manage risks and conduct assessments is of the utmost importance. Still, the act does not go into details on the exact expectations regarding risk management, which leaves a lot of room for interpretation. It further does not provide any details on solutions or even guidelines for implementation. Kumar et al. [4] revealed in an industry survey that a knowledge gap exists when it comes to securing machine learning systems. The US National Security Commission on Artificial Intelligence [5] highlights in their final report that it is critical for the success of AI systems to be trustworthy and to work as intended (i.e. in a predictable way). Furthermore, it is stated that AI systems will become a more popular target of attacks.

The major contribution of this paper is the introduction of an AI Risk Management System, which can be used to systematically identify, assess, treat and monitor risks in the artificial intelligence domain. The focus lies on the introduction of a general risk management framework for AI. Therefore, the paper does not cover specific tools and techniques in detail, which can be used to implement the framework in a specific environment.

The remainder of this paper is structured as follows: Section II surveys relevant literature in the domain of risk management and AI security. Furthermore, it outlines relevant requirements for risk management systems. Following, Section III highlights the requirements, we analyzed with regards to the proposed AI act. In the succeeding section, we introduce our risk man- agement system approach for artificial intelligence (AIRMan), before we draw our conclusions and highlight future work in Section V.

979-8-3503-2437-2/22/$31.00 ©2022 IEEE
DOI 10.1109/AEIS59450.2022.00017

II. Requirements

There is a vast amount of literature, which focuses on information security risk management or risk management in general. In this subsection, we provide an overview on general risk management standards and best practices as well as current developments in the field of AI risk management approaches.

One of the most well-known and widely adopted standards in the risk management domain is the international standard ”ISO 31000:2018 Risk Management” [6]. It outlines three main aspects: Firstly, it highlights principles for effective and efficient risk management. Secondly, it introduces a framework to support the integration of risk management into business operations and activities. Thirdly, it provides details on the iterative risk management process.

ISO 31000 establishes eight principles, which are essential for effective and efficient risk management [6]:

- Integrated [6]: Risk management should be an important part of organizational activities.
- Structured and comprehensive [6]: Risk management should take a structured and holistic approach to achieve comparable and consistent results.
- Customized [6]: Risk management framework should be appropriate for the organization and tailored to its characteristics and objectives.
- Inclusive [6]: Risk management should timely involve stakeholders to enable the incorporation of their knowledge, views and perceptions.
- Dynamic [6]: Risk management should be able to react on dynamic changes in the risk situation of an organization.
- Best available information [6]: Risk management inputs should be derived information about the past, the present and the future.
- Human and cultural factors [6]: Risk management should take into account human and cultural aspects, as these have a significant impact on risk management.
- Continual improvement [6] [7]: Risk management should learn over time and carry out continuous improvement efforts. All management systems (cf. [8]) and most enterprise-wide used risk management approaches highlight the importance to continual improve performance. Therefore, it is key to integrate a feedback mechanism to facilitate improvement over time.

In order to harmonize the management standards structure and enable the operation of integrated management systems, ISO published in its Annex SL - ISO Management standards [8] a general structure (i.e. Scope, Normative references, Terms and definitions, Context of the organization, Leadership, Planning, Support, Operation, Performance Evaluation, Improvement), which lays the foundation for ISO management systems (e.g. ISO 9001, ISO 20000-1, ISO 22301, ISO 27001). Furthermore in Appendix 2 of Annex SL [9] guidance for authors of management systems is provided. Therefore, Annex SL served as a valuable input to structure and elaborate our AI risk management system.

In order to support practitioners with tools and techniques, ”ISO/IEC 31010:2019 - Risk Management Techniques” [10] details commonly used approaches in the industry.

The standard ISO 27005 [7], [11] provides detailed information on how the framework and process - described in ISO 31000 [6] - can be implemented in the information security domain. Furthermore, ISO 27005 [7] outlines information on various risk management components, such as example risk criteria, example threats or example vulnerabilities.

Another well-known approach is COBIT 5 for Risk [12]. The guideline highlights three types of risk categories (i.e. IT Benefit/Value Enablement, IT Programme and Project Delivery, IT Operations and Service Delivery). The guide outlines essential risk management approaches and information required to develop the risk profile of a company. In order to describe the risks (i.e. the risk scenarios) the elements *Actor, Threat Type, Event, Asset/Resource, Time are used*. For risk mitigation, COBIT for Risk makes use of the seven COBIT 5 enablers *Principles, Policies and Frameworks; Process; Organizational Structures; Culture, Ethics and Behaviour; Information; Services, Infrastructure and Applications;People, Skills and Competencies*. COBIT provides a publication on 111 risk scenario examples [13].

While there exist various other approaches for risk management (e.g. OCTAVE [14], NIST SP 800-30 [15]), most of them are not tailored to artificial intelligence but support the general risk management process.

One of the few frameworks, which specifically deals with risk management for AI is the recently proposed NIST AI Risk Management Framework concept paper [16]. The proposed structure consists of the following three components:

- *Core components* are used to outline activities and deliverables required to manage AI risks.
- *Profile components* are used for prioritization of AI tasks and offer a way to incorporate context-specific information.
- *Implementation Tier components* support the realization of AI risk management by providing decision making and communication support.

For its AI Risk Management Framework initiative, NIST recently developed a draft taxonomy for risks [17] aiming to facilitate the management of risks for stakeholders. The three broad risk categories, which have been identified are:

- Technical design attributes (i.e. aspects under control of the system designer/developer)
- How AI systems are perceived (i.e. mental representations of models)
- Guiding policies and principles (i.e. broader societal determinations of value)

For each of the three categories, the publication maps important characteristics [17], which can be a risk source if not adequately present. Furthermore, in [17] the authors map their risk categories and taxonomy to publications on Trustworthy AI of relevant policy makers (i.e. EU [18], OECD [19], US [20]).

Cheatham et al. [21] highlight five issues that give rise to AI risks (i.e. data difficulties, technology troubles, security snags, models misbehaving, interaction issues). Furthermore, the authors identified 13 sample impact categories, which they grouped into the three categories Individuals (i.e. Physical safety, privacy & reputation, digital safety, financial health, equipment & fair treatment), Organizations (financial performance, non-financial performance, legal & compliance, reputational integrity) and Society (i.e. national security, economic stability, political stability, infrastructure integrity). Derived from their analysis, the publication concluded three core principles for AI risk management (i.e. structured risk identification approach, implementation of robust enterprise-wide controls, reinforcement of specific controls depending on the risk nature) [21].

Another issue was raised by Holzinger et al. [22]: For all the benefits of AI, the large-scale and ubiquitous adoption of AI holds enormous and unimagined potential for novel, unforeseen threats. Therefore, all stakeholders involved in the development process must ensure that AI is developed with these potential threats in mind. In the future, it will be essential that the security, traceability, transparency, explainability, validity, and verifiability of AI applications are ensured in everyday routine operations especially when AI components are used as standard tools by non-professionals.

This also requires the integration of ethical [23] issues as the explainability problem also results in decision makers needing to rely on the results of AI components to be free of bias and taking care of other ethical problems. This also holds true for legal aspects of using AI in sensitive environments [24] to ensure the use of trustworthy and ethically reliable AI and to help avoid the misuse of AI.

One aspect that needs additional attention in the risk management process is caused by the heavy reliance on data, especially in cases the source of this data is outside of the AI system operators control: Applications, industrial as well as commodity software, typically produce a certain amount of erroneous data, thus requiring data cleansing and preparation steps before the data can actually be used, thus making changes to the source data. The legal implications of changes to the resulting models and the distortion in subsequent AI results pose another risk that requires additional research [25].

The ENISA threat landscape report [26] outlines current cyber security challenges concerning artificial intelligence, which can be used to derive risks. It maps the 74 threats to an introduced generic AI lifecycle and relevant AI assets, which have been grouped into six categories (Data, Model, Actors, Processes, Environment/Tools, Artefacts).

Another good source for better understanding the threat landscape of artificial intelligence is MITRE's ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) matrix [27].

It outlines the tactics and techniques used by an adversary throughout the various stages of an attack (i.e. Reconnaissance; Resource Development; Initial Access; ML Model Access; Execution; Persistence; Defense Evasion; Discovery; Collection; ML Attack Staging; Exfiltration; Impact). The modeling of the attacks follows the structure of MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge) matrix [28].

Another approach is outlined by the World Economic Forum's AI risk assessment toolkit [29] as part of their guidance on AI procurement. Therein several yes/no questions are provided to the determine the risk situation with regards to AI procurement.

III. Risk Management wihin the AI Act

Due to the numerous research strands and fast developments of the recent past, the definition of artificial intelligence is a challenging task.

The European Commission's proposal for an AI Act [3] defines the "artificial intelligence system (AI system)" very broadly to ensure that the definition is future-proof [30].

However, the main bulk of the regulation focuses on a specific subset of AI systems: the so called "High-Risk AI Systems". While the definition of AI systems is very broad, the definition of "High-Risk AI system" is narrowed down in Art. 6 of the AI Act to certain objectives, that are more likely to result in a high risk for a person (e.g. by misuse or for manipulative, exploitative and social control practices).

These risks are addressed within the AI Act [3] by mandatory requirements that high-risk AI systems have to comply with.

According to Rec. 27 AI Act, these requirements should ensure, that high-risk AI systems do not pose unacceptable risks to important Union public interests as recognised and protected by Union law.

These requirements include i.a. data governance (Art. 10), technical documentation (Art. 11), record-keeping (Art. 12), transparency and provisioning of information to users (Art. 13), human oversight (Art. 14), accuracy, robustness and cybersecurity (Art. 15) – all of which should apply, "taking into account the intended purpose of the use of the system and according to the risk management system to be established by the provider" (Rec. 42 AI Act). This is the reason why the risk management can be found as the first material requirement in Art. 9 AI Act.

Art. 9 of the AI Act states that a **risk management system** shall be "established, implemented, documented and maintained in relation to high risk AI systems".

Furthermore, it shall consist of a "continuous iterative process run throughout the entire life cycle of a high-risk AI system, requiring regular systematic updating. [. . .]". In this regard the risk management system as defined in the AI Act is referring to common definitions of risk management.

For credit institutions, this risk management will be part of the established risk management procedures according to the credit institutions directive.

The key characteristics of a risk management system that addresses risks of AI systems are highlighted in the following obligatory steps that it shall comprise (Art. 9(2) AI Act):

a) identification and analysis of the known and foreseeable risks associated with each high-risk AI system;
b) estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose and under conditions of reasonably foreseeable misuse;
c) evaluation of other possibly arising risks based on the analysis of data gathered from the post-market monitoring system referred to in Art. 61;
d) adoption of suitable risk management measures in accordance with the provisions of the following paragraphs.

Art. 9(3) AI Act states that such suitable risk management measures shall give due consideration to the effects and possible interactions resulting from the combined application of the requirements in that chapter.

It is also important to note that Art. 9(3) specifically states that risk management measures shall take into account the generally acknowledged state of the art, including as reflected in relevant harmonised standards or common specifications.

However, not every risk associated with high-risk AI systems has to be eliminated, as long as the overall residual risk situation of the high-risk AI systems is judged acceptable.

Residual risks have to be documented (Art. 11 and Annex IV(3) and (4) AI Act) and communicated to the user.

To identify the most appropriate risk management measures, Art. 9(4) AI Act requires the following:

a) elimination or reduction of risks as far as possible through adequate design and development;
b) where appropriate, implementation of adequate mitigation and control measures in relation to risks that cannot be eliminated;
c) provision of adequate information pursuant to Art. 13 (in particular if used with its intended purpose and reasonably foreseeable misuse) and, where appropriate, training to users.

This requires an evaluation of the technical knowledge, experience, education, training to be expected by the user and the environment in which the system is intended to be used.

Specific consideration should also be given to whether the high-risk AI system is likely to be accessed by or have an impact on children (Art. 9(8) AI Act).

Art. 9(5)-(7) set further requirements to the **testing of high-risk AI systems**, which shall be performed, as appropriate, at any point in time throughout the development process, and, in any event, prior to the placing on the market or the putting into service.

According to these articles, testing should serve the following purposes:

1) identifying the most appropriate risk management measures
2) ensure that high-risk AI systems perform consistently for their intended purpose
3) ensure that high-risk AI systems are in compliance with the requirements of (Chapter II of) the AI Act.

Testing shall be made against defined metrics and probabilistic thresholds that are appropriate to the intended purpose and should not go beyond of what is necessary to achieve the intended purpose of the AI system.

As mentioned before, these risk management measures are the basis for compliance with other requirements for high-risk AI systems, as the appropriate measures to be taken within these requirements depend on the risk assessment. All these requirements are of course intertwined, as risk management is a continuous iterative process run throughout the entire life cycle of the AI system.

Art. 10 further requires **data governance and management** practices with regard to data sets used for training, validation and testing. As training, validation and testing data sets shall be relevant, representative, free of errors and complete (Art. 10(3) AI Act) this requires a risk assessment according to Art. 9(2)(a)-(c), and the choosing of the correct data set that addresses these risks (e.g. biases, restricted functioning for a specific setting) is part of the adoption of suitable risk management measures (Art. 9(2)(d) AI Act).

The **technical documentation** under Art. 11 has to be drawn up in such a way to demonstrate that the high-risk AI system complies with the requirements, which includes a detailed description of the risk management system under to Art. 9 AI Act (see Annex IV(4) AI Act).

Record-keeping as defined in Art. 12 AI Act requires logging capabilities, that ensure a level of traceability of the AI system's functioning throughout its life cycle that is appropriate to the intended purpose of the system. This is important for compliance with Art. 9 as it allows a continuous iterative application of the risk management system (as well as for the post-market monitoring, Art. 61 AI Act).

Risk management is also required to provide information according to Art. 13 AI Act (e.g. the "characteristics, capabilities and limitation of performance") as well as to design human-machine interface tools that allow for effective human oversight (Art. 14 AI Act) or to define the necessary requirements for accuracy, robustness and cybersecurity (Art. 15 AI Act).

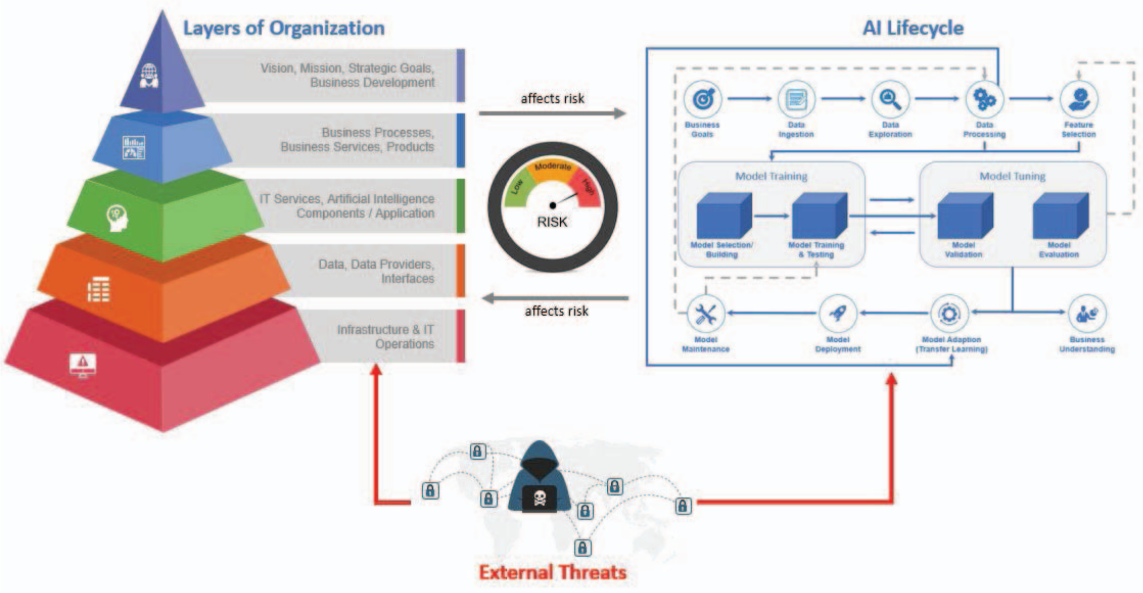

Fig. 1. Risk affecting the AI Lifecycle [3]

Due to the penalties under the AI Act the non-compliance with its obligations can have a severe financial impact on the offender, as the administrative fines can go up to € 20.000.000,– or if the offender is a company, up to 4% of its total worldwide annual turnover for the preceding financial year, whichever is higher. In case of non-compliance with the prohibitions under Art. 5 or with the obligation of the above mentioned Art. 10 AI Act, these fines are even up to € 30.000.000,– or 6% of the worldwide annual turnover; the supply of incorrect, incomplete or misleading information to notified bodies/authorities is also subject to severe fines (€ 10.000.000,–; 2% worldwide annual turnover).

The risk management system under Art. 9 AI Act is not only as itself a requirement and an obligation in the sense of Art. 71(4) but also the logical basis for the fulfilment of the other requirements, as shown above (including the requirement of Art. 10 AI Act) which means that under the AI Act the appliance of an effective risk management system is an essential prerequisite.

IV. AI Risk Management System (AIRMan)

In this section, we introduce our AI Risk Management System (AIRMan), which we derived and aligned from the structure of existing management systems literature (e.g. ISO Annex SL or ISO 31000), as outlined in Section II as well as with the requirements of the Proposal of the AI Act (see Section III).

The general structure aims at facilitating continuous improvement using the widely applied PDCA (Plan-Do-Check-Act) cycle [31].

The adherence to ISO Annex SL is essential as it facilities the integration into other (existing) management systems as demanded in Art. 17 of the AI Act [3].

Derived from the basic structure of Annex SL [8] and influenced by COBIT5/2019 [32], [33] and ISO31000 [6], we came up with the following structure.

A. Plan

This phase starts with gathering a deep understanding of the organization, its products and its stakeholders.

The results of this analysis are critical to establish clear goals and strategic objectives for management systems, to determine the scope and define a systematic approach to manage risks as well as relevant criteria (e.g. risk evaluation criteria, risk acceptance criteria, ...).

For ensuring a smooth integration into other management systems (e.g. Information Security Management System, Business Continuity Management System, Privacy Information Management System, Quality Management System) it is essential to determine appropriate interfaces and reporting structures between these systems.

As leadership is a key success factor, management has to demonstrate management commitment. Examples for activities highlighting management commitment comprise:

- Communicate the importance of trustworthy AI and AI risk management activities, especially for high-risk applications
- Establish an AI risk management policy, which ensures that the AI risk management approach and management system is aligned with the organizational objectives and compliant to the AI Act
- Determine and sign-off the risk appetite for AI applications in alignment of the organization's risk acceptance criteria

- Allocate adequate financial and human resources for risk analysis and mitigation of artificial intelligence applications
- Actively participate in meetings where risks of AI applications are discussed
- Review the performance of the AI risk management system and set corrective actions when required.

An important artefact of this phase is the AI Risk Management Policy, which sets the cornerstones for the management system and communicates the importance of the topic within the organization.

Relevant content of this policy comprises the following items:

- Definition of relevant terms, e.g. artificial intelligence, ...
- Importance of security and risk management for AI applications
- Clear statement, which demands the compliance with applicable laws (e.g. AI Act, General Data Protection Regulation)
- Rules for categorization of AI systems, e.g. using the proposed structure of [3], [34] (unacceptable, high risk, limited risk, minimal risk)
- Roles and responsibilities for AI risk management
- Systematic approach to risk identification, risk assessment and risk mitigation

Another essential aspect of the phase is the provisioning of resources. Besides the need for financial resources, it is critical to have skilled personnel for performing the tasks within the management system and create an adequate risk and security culture amongst the employees of the organization.

In order to specify the relevant skill set, well-known competence frameworks (e.g. NICE workforce [35], EDISON [36], Skills Framework for the Information Age [37]) covering the topics security, risk and artificial intelligence can be used.

As correctness and traceability are important characteristics of management systems, it is important that adequate documentation takes place.

While the degree of documentation can vary by size, requirements and type of organization, documentation should at least include:

- AI risk management policy & scope
- Risk assessments & business impact analyses
- Risk treatment strategies & plans
- Audits and performance measurements of controls
- Management reviews
- Lessons learned and corrective measures taken.

B. Do

The Do-Phase is the core phase of every management system. Before discussing the individual steps of this phase in more detail, it is important to highlight the interdependencies between the layers of the organizations and the AI lifecycle. Figure 1 provides a short overview.

The AI lifecycle spans across all levels of an organization. Therefore, as outlined in Figure 1 all layers within the enterprise architecture can on the one side contribute to the risks within the AI lifecycle and on the other side can potentially be influenced by the arising threats caused by AI development and operations.

On the top layer, the strategic goals and envisioned business development can have a long-term effect on AI risks. Making risk management and security a priority can have positive effects, while neglecting the importance of this topic can lead to high risks that won't be treated in an appropriate manner.

On business processes level it is important that business process owners provide sufficient information about the possible impacts that could arise if risks materialize within the AI lifecycle. Furthermore, security requirements and risk acceptance have to be determined at this level.

On the Application and IT service layer, risks of used components play a vital role. Errors in libraries or changes in software/AI components can lead to new threats that can be hard to detect in unstructured development environments.

Data stored with external data providers and interfaces to data services play a major role in the development, implementation and maintenance of AI systems. Therefore, occurring risks on this layer may lead to severe consequences.

The bottom layer (Infrastructure & IT Operations) is responsible for the secure and resilient operation of AI applications. Compromised systems or failure within the infrastructure or logging can lead to performance degradation, wrong outputs or service disruptions.

It is also important to note that internal and external threats can arise on each layer and therefore the AI risk management must not be performed isolated by AI developers, architects and engineers.

An overview on sample risks affecting the different layers can be found in Table I.

TABLE I. SAMPLE RISKS AFFECTING THE DIFFERENT ORGANIZATIONAL LAYERS

Layer	Sample Risks
Vision, Strategic Goals	Compliance risks, reputation risks ...
Business Process, Products	Service delivery risks, malfunction of product, process interruption …
IT Services, AI Components	Adversarial machine learning risks …
Data, Data Providers	Integrity risks, confidentiality/privacy risks …
Infrastructure, IT Operations	Availability risks …

Additionally, in the following paragraphs, we briefly outline sample risks, which might occur in the individual phases of the AI lifecycle (see Figure 2).

- Business Goals: Data needed to answer business questions has not yet been collected and is not available.

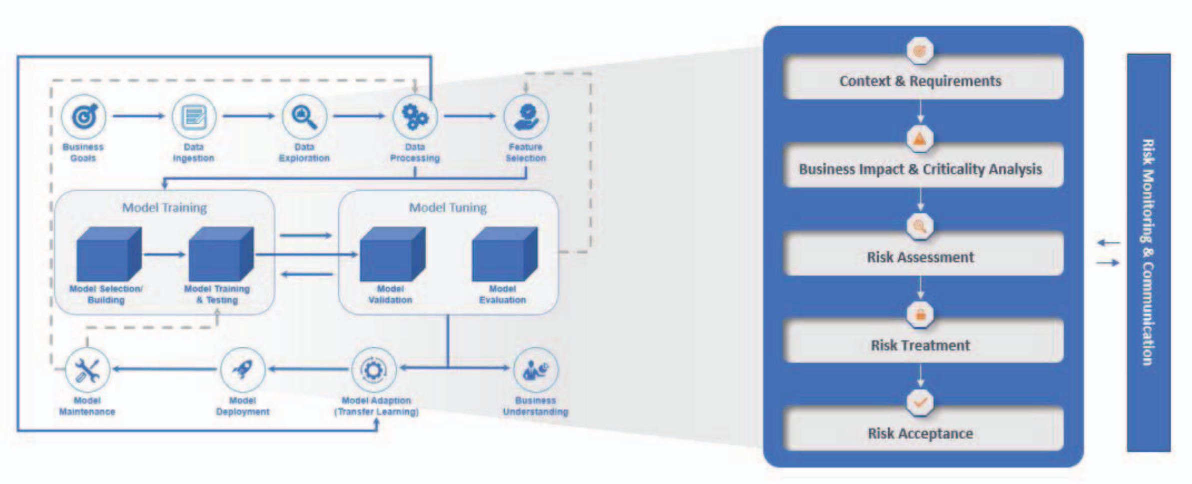

Fig. 2. Proposed risk management process derived from [6], [11], [38] for AI Lifecycle [3]

- Data Ingestion: Data is stored in separate locations in form of data silos and prevents a holistic view of organisation data.
- Data Exploration: A machine learning model is chosen without first examining the distribution of its data. Promptly, the model is chosen that requires normally distributed data, even though the data is not normally distributed. Additionally unbalanced (biased) data is not recognised due to lack of data visualisation.
- Data Processing: Incomplete data or non-transformed data lead to incorrect results of a machine learning model.
- Feature Selection: All available data is passed unfiltered to a model and the necessity of meaningful features is ignored. Even irrelevant features are not cleaned in advance.
- Model Training: A deep learning network is the only model trained on the data, the evaluation process is done without comparison to a simple baseline model. A good failure analysis of models does not take place.
- Model Tuning: The Model Tuning does not take place together with Model Training, it is waived. The hyper parameters of models are not optimized, neither manually nor automated.
- Model Deployment: The path from data collection to data processing to the model is not well documented and cannot be traced if problems arise during operation.
- Model Adaptation: A model trained and optimized for a specific task is transferred to another similar area without taking into account that there must be adaptations of the model. Using a pre-trained model is just a starting point and must be further trained, tuned and evaluated for the new application.
- Model Maintenance: Regular updating after the deployment of the model was not considered, although external influences like sensors, age or the environment change regularly.

In the following, we highlight the phases of AI risk management, which have been influenced by ISO 31000 [6], ISO 27005 [11], ISO 22301 [38], NIST SP800-30 [15] and COBIT for Risk [12].

Context and requirements: The first step is the detailed specification of the scope of the analysis. This includes the determination of business activities and assets which should be further analyzed.

Business Impact and Criticality Analysis: This step aims at determining the importance of AI components and highlights the consequences on business, which could arise (over time) if the components are not working as intended. Considering the criticality and consequences over time enables a solid estimation/assessment of the importance of AI components without having to consider each individual risk.

Risk Assessment: Risk assessment paves the way to determine risk treatment options adequate for the risk situation of the organization. This step starts with the identification of risks. Risk identification is a crucial task as further risk management activities cannot consider what has not been identified. This can be performed by either interviews or risk/hazard lists & scenarios. Good resources for risk identification include MITRE ATLAS [27], ENISA's AI Cybersecurity Challenges [26] or KPMG's AI Risk and Control Matrix [39].

After potential risks have been identified, the risks have to be analyzed in more detail in order to enable a prioritization. In order to facilitate the analysis, different parameters, such as impact of risks on assets (e.g. processes, services, information) or probability of occurrence, have to be determined.

Depending on the selected approach, also further parameters (e.g. frequency, detectability) can be used. As a result, risks can be prioritized and a selection of risk treatments can take place in the succeeding step.

Risk Treatment: The four key strategies to address risks are risk acceptance, risk mitigation, risk transfer and risk avoidance. Depending on the determined severity of the risk according to the previous analysis, one of the strategies is applied. The most common strategy is risk mitigation leading to the definition and implementation of security controls.

Risk Communication: Risk communication is key to gather risk information from relevant stakeholders and to provide them with information about risks, which facilitate well- informed decisions about risk handling strategies and to raise awareness.

Risk Monitoring: Risks in the digital area are often very dynamic, as technologies evolve quickly, vulnerabilities can be exploited at a fast pace and are not restricted by physical borders. Furthermore, due to the complexity of AI systems, risks associated with unexpected changes of the environment or overlooked errors in the learned models considerably increase the risk situation. Therefore, it is critical to implement an adequate risk monitoring system. Depending on the criticality of the risk, the AI approach used and the dynamics of the environment, the monitoring frequency should be carried out at periodic intervals, at major changes or in a continuous manner.

Actually, the results of a machine learning model should be reviewed continuously (this is maybe not always possible). Changes in data collection or data processing can directly affect models, any update of infrastructure can lead to unforeseen effects. Thus, it is important to plan such changes early to adapt models at the same time.

Situations in which a model has to learn new behavior quickly leads to a continuous adaptation of the model, because otherwise it would not react well. An example would be the automatic detection of cyber-attacks, which adapt almost daily. Additionally attacks on machine learning algorithms are likely to increase in the near future. Such attacks are made to bring models to unexpected output. It is important to recognize such changes which only happen through vigilance and continuous checking.

C. *Check*

The main objective of this phase is to ensure the effectiveness of the management system. As a result non-conformities and opportunities for improvement are highlighted, which are the basis for continual improvement.

Essential tasks of this phase comprise, amongst others, monitoring of AI risk management system activities, (internal & external) audits/assessments and management reviews.

Special attention has to be put on the issue of audits and assessments in this phase: External reviewers face a challenging task when being required to analyze the completeness of AI risk management and the effectiveness of applied countermeasures.

Furthermore, even the application of certain technical measures, e.g. special attention to bias in source data, is almost impossible to verify when only being able to check on the trained model - re-training for comparison, on the other hand, can be extremely costly and time-consuming, thus making it a practically impossible strategy to follow.

The same holds true in case of shared models that have been pre-trained by external entities. Depending on the actual application, sanity checks could be put into place in order to audit the system. Compared to traditional software, some AI techniques constantly change their model when being presented with new data, thus making standard practices from the security world, like fuzzying, problematic, as the random input might introduce issues (like bias) that have not been present in the original system.

D. *Act*

The last phase of the management system lifecycle is dedicated to the continual improvement. Revealed deficiencies are eliminated through the implementation of corrective actions and improvement opportunities realized.

This might require a lot of management effort in case of relying on external resources, especially when considering pretrained models or external data sources. In addition, due to the explainability problem, verification in this phase might be difficult to achieve.

V. Conclusion & Future Work

It is undeniable that the usage of AI applications leads to a variety of benefits. At the same time, as a result of this development, systems get more complex and organizations become more dependent on trustworthy AI. Therefore, risks associated with AI must be addressed appropriately.

With the proposal of the AI Act, the European Union made an important step towards regulating high risk AI systems (e.g. AI used in safety components). A key requirement in the proposal is the operation of a risk management system.

In this paper, we firstly analyzed relevant literature in the area of risk management, management systems and AI risk as well as the relevant components of the AI Act. Thereafter we introduced the AIRMan approach, which aims at facilitating affected parties to tackle the challenges arising with the upcoming need to comply to the proposed regulation.

The presented management system supports organizations independent from their industry to implement a systematic approach to identify, analyze, monitor and treat AI risks.

In our future work we plan to tailor existing risk assessment tools and techniques to fit the special demands of AI systems. Furthermore, we plan to elaborate generic AI controls for common AI risks.

We will further focus our research on way to audit AI controls, despite the challenges posed by complexity and non-explainability.

Acknowledgment

This research was funded by the Austrian Research Promotion Agency (FFG) through COIN project 883977 “Secure Supply Chains for Critical Systems” and the APOG project. The financial support by the Austrian Research Promotion Agency, the Federal Ministry for Digital and Economic Affairs and the state of Lower Austria is gratefully acknowledged.

REFERENCES

[1] A. Holzinger, M. Dehmer, F. Emmert-Streib, R. Cucchiara, I. Augenstein, J. Del Ser, W. Samek, I. Jurisica, and N. D´ıaz-Rodrıguez, "Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence," Information Fusion, vol. 79, no. 3, pp. 263–278, 2022.

[2] S. Tjoa, C. Buttinger, K. Holzinger, and P. Kieseberg, "Penetration testing artificial intelligence." ERCIM News, vol. 2020, no. 123, 2020.

[3] European Commission, "Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts," https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206, accessed December 2021.

[4] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning - industry perspectives," in 2020 IEEE Security and Privacy Workshops (SPW). IEEE, 2020, pp. 69–75.

[5] E. Schmidt, R. Work, S. Catz, E. Horvitz, S. Chien, A. Jassy, M. Clyburn, G. Louie, C. Darby, W. Mark, K. Ford, J. Matheny, J.-M. Griffiths, K. Mcfarland, and A. Moore, "National security commission on artificial intelligence (ai)," National Security Commission on Artificial Intelligence, Tech. Rep., 2021.

[6] International Organization for Standardization, "ISO 31000:2018 - Risk management — Guidelines," 2018.

[7] ——, "ISO/IEC DIS 27005:2021 Information security, cybersecurity and privacy protection — Guidance on managing information security risks," 2021.

[8] ——, "Annex SL (normative) Harmonized approach for management system standards," https://www.iso.org/sites/directives/current/consolidated/index.xhtml, 2021, accessed December 2021.

[9] ——, "Annex SL - Appendix 2 (normative) Harmonized structure for MSS with guidance for use ," https://isotc.iso.org/livelink/livelink/fetch/-8921878/8921901/16347356/16347818/2021-05 Annex SL Appendix 2.pdf?nodeid=21826538&vernum=-2, 2021, accessed December 2021.

[10] ——, "ISO/IEC 31010:2019 - Risk management — Risk assessment techniques," 2019.

[11] ——, "ISO/IEC 27005:2018 Information technology — Security techniques — Information security risk management," 2018.

[12] ISACA, "COBIT 5 for Risk," 2013.

[13] ——, "Risk Scenarios Using COBIT 5 for Risk," 2014.

[14] C. Alberts, A. Dorofee, J. Stevens, and C. Woody, "Introduction to the octave approach," Carnegie-Mellon University, Software Engineering Institute, Tech. Rep., 2003.

[15] National Institute of Standards and Technology (NIST), "NIST Special Publication 800-30: Guide for ConductingRisk Assessments rev. 1," https://csrc.nist.gov/publications/detail/sp/800-30/rev-1/final, Sep. 2012.

[16] National Institute of Standards and Technologies (NIST), "AI Risk Management Framework Concept Paper," https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper 13Dec2021 posted.pdf, Dec. 2021, accessed December, 2021.

[17] National Institute of Standards and Technology (NIST), "Draft - taxonomy of ai risk," https://www.nist.gov/system/files/documents/2021/10/15/taxonomy AI risks.pdf, Oct. 2021, accessed December 2021.

[18] European Union, "Ethics guidelines for trustworthy AI," https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019.

[19] Organisation for Economic Co-operation and Development - OECD, "AI Principles," https://oecd.ai/en/ai-principles.

[20] "Executive Order 13960 - Promoting the Use of Trustworthy Artificial Intelligence in the Federal Government," https://www.federalregister.gov/documents/2020/12/08/2020-27065/promoting-the-use-of-trustworthy-artificial-intelligence-in-the-federal-government.

[21] B. Cheatham, K. Javanmardian, and H. Samandari, "Confronting the risks of artificial intelligence," McKinsey, Tech. Rep., 2019. [Online]. Available: https://www.mckinsey.com/~/media/McKinsey/BusinessFunctions/McKinseyAnalytics/OurInsights/Confrontingtherisksofartificialintelligence/Confronting-the-risks-of-artificial-intelligence-vF.pdf

[22] A. Holzinger, E. Weippl, A. M. Tjoa, and P. Kieseberg, "Digital transformation for sustainable development goals (sdgs) - a security, safety and privacy perspective on ai," in Springer Lecture Notes in Computer Science, LNCS 12844. Cham: Springer, 2021, pp. 1–20.

[23] H. Mueller, M. T. Mayrhofer, E.-B. V. Veen, and A. Holzinger, "The ten commandments of ethical medical ai," IEEE COMPUTER, vol. 54, no. 7, pp. 119–123, 2021.

[24] D. Schneeberger, K. Stoeger, and A. Holzinger, "The european legal framework for medical ai," in International Cross-Domain Conference for Machine Learning and Knowledge Extraction, Springer LNCS 12279. Cham: Springer, 2020, pp. 209–226.

[25] K. Stoeger, D. Schneeberger, P. Kieseberg, and A. Holzinger, "Legal aspects of data cleansing in medical ai," Computer Law and Security Review, vol. 42, p. 105587, 2021.

[26] European Union Agency for Cybersecurity (ENISA), "AI Cybersecurity Challenges - Threat Landscape for Artificial Intelligence," https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges, Dec. 2020, accessed: December 2021.

[27] MITRE, "MITRE-ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)," https://atlas.mitre.org/, accessed November 2021.

[28] ——, "MITRE-ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge)," https://attack.mitre.org/, accessed November 2021.

[29] World Economic Forum, "AI Procurement in a Box: AI Government Procurement Guidelines," https://www3.weforum.org/docs/WEFAI Procurement in a Box Workbook 2020.pdf, Jun. 2020.

[30] J. Zanol, A. Buchelt, S. Tjoa, and P. Kieseberg, "What is "AI"? Exploring the scope of the "Artificial Intelligence Act"," in Data Governance & Privacy: Tagungsband des 25. Internationalen Rechtsinformatik Symposions IRIS 2022/Proceedings of the 25th International Legal Informatics Symposium IRIS 2022. Editions Weblaw, 2022, pp. 75–82.

[31] M. Sokovic, D. Pavletic, and K. K. Pipan, "Quality improvement methodologies–pdca cycle, radar matrix, dmaic and dfss," Journal of achievements in materials and manufacturing engineering, vol. 43, no. 1, pp. 476–483, 2010.

[32] ISACA, "COBIT 2019 Framework - Introduction and Methodology," 2018.

[33] ——, "COBIT 2019 Framework - Governance and Management Objectives," 2018.

[34] European Commission, "Regulatory framework proposal on artificial intelligence," https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai, Oct. 2021, accessed January 2022.

[35] R. Petersen, D. Santos, M. C. Smith, K. A. Wetzel, and G. Witte, "NIST SP800-181 rev. 1 Workforce Frameworkfor Cybersecurity (NICE Framework) ," https://doi.org/10.6028/NIST.SP.800-181r1, Nov. 2020.

[36] "EDISON Data Science Framework," https://edison-project.eu/edison/edison-data-science-framework-edsf/, accessed December 2021.

[37] "Skills Framework for the Information Age," https://sfia-online.org/en/sfia-8, accessed January 2022.

[38] International Organization for Standardization - ISO, “ISO 22301:2019 Security and resilience — Business continuity management systems — Requirements,” https://www.iso.org/obp/ui#iso:std:iso:22301:ed-2:v1:en, 2019.

[39] KPMG, “AI Risk and Controls Matrix,” https://assets.kpmg/content/dam/kpmg/uk/pdf/2018/09/ai-risk-and-controls-matrix.pdf, 2018