CyCon 2024: Over the Horizon
16th International Conference on Cyber Conflict
C. Kwan, L. Lindström, D. Giovannelli, K. Podiņš, D. Štrucl (Eds.)

Artificial Intelligence System Risk Management Methodology Based on Generalized Blueprints

Dan Bogdanov*
Information Security Research Institute
Cybernetica, Estonia

Paula Etti*
Information Security Research Institute
Cybernetica, Estonia

Liina Kamm*
Information Security Research Institute
Cybernetica, Estonia
liina.kamm@cyber.ee

Fedor Stomakhin*
Information Security Research Institute
Cybernetica, Estonia

Abstract: The rapid uptake of artificial intelligence (AI) systems requires similar advances in their governance. Public and private sector institutions want to adopt new AI tools as they perceive potential efficiency gains and value from them. As with every technological advance, the uptake phase of AI is the ideal time to improve the governance, cybersecurity and safety of these systems.

The cybersecurity risks in AI systems are similar to the ones in other information technology systems. However, the regulation of AI systems is changing, so new governance tools are needed. Furthermore, the safety and societal impact of AI depends on the technological choices made when building the systems (e.g., biased training data, overfitted machine learning models, model poisoning attacks or needlessly computation-heavy algorithms).

AI tools built with large language model technology seem to speak our languages and therefore appear deceptively easy to adopt. The goal of our research is to provide risk management tools that are similarly easy to use, even if they later lead the adopter into setting up a full technical quality management system.

We have created three blueprints of AI system deployments to which an organization deploying AI can match their use case. For each blueprint, we have created high-

* Co-funded by the European Union. The views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Cybersecurity Competence Centre. Neither the European Union nor the European Cybersecurity Competence Centre can be held responsible for them. This work was supported by the Estonian Centre of Excellence in AI (EXAI), funded by the Estonian Ministry of Education and Research.

level guidance on which cybersecurity, data rights and ethical aspects the deploying organization needs to consider. Those building AI systems can quickly match their use cases against the blueprints and speed up the secure and ethical adoption of AI.

Keywords: *artificial intelligence, cybersecurity, data governance, data protection, risk management*

1. INTRODUCTION

Artificial intelligence (AI) is developing at a rapid pace and organizations are deploying AI systems to make their processes more efficient. However, the deployment of AI systems brings with it the need for risk management to ensure the deployed systems are secure, safe and compliant with all relevant laws. This can be a difficult task for smaller organizations that do not have a mature risk management framework in place. Even large organizations with an established risk management process must account for AI-specific risks.

In this paper, we consider AI systems that are IT systems. Several frameworks exist for cybersecurity risk management in IT systems and our goal is to simplify their adoption. We propose three blueprints that cover common deployments of AI systems and present a methodology to manage the risks based on these blueprints. Our approach is compatible with existing methodologies and can be either integrated with existing information security management systems or be used to start a new one.

We take into account cybersecurity, regulatory and AI-specific risks. As the cybersecurity risks in AI systems are similar to those in other IT systems, we focus here on AI-specific and regulatory risks. Our goal with this paper is to encourage more organizations to achieve basic AI security and safety.

2. AI IN THE CONTEXT OF SYSTEMS

Artificial intelligence has a variety of definitions in regulation and standards. Given the rapid pace of AI technology development, some definitions have become obsolete and need to evolve. The challenge lies in striking a balance between specificity and flexibility, ensuring that definitions are robust enough to guide current applications and future innovations.

The ISO/IEC 22989:2022 standard defines an AI system as an engineered system generating an output for a given set of human-defined objectives [1]. The definition in the draft European AI Act (AIA) [2] is similar but adds an (extendable) list of techniques and stresses interaction with its environment. The Organization for Economic Co-operation and Development (OECD) definition is similar to both but considers inputs and the autonomy of AI systems [3]. The United States' AI Bill of Rights uses the broader term automated systems [4].

A. Stakeholders of AI Systems

The listed sources also describe AI stakeholders. The ISO/IEC 22989:2022 standard provides a hierarchy of roles relevant to AI systems (see Section 5.19 and Figure 2 in [1]). The standard differentiates between AI providers, producers, customers, partners, subjects and relevant authorities. Such a differentiation is very helpful in discussing AI system deployments and their cybersecurity.

At the same time, for regulatory and legal discussions, definitions in regulation are more relevant. For example, the AIA definitions are more tailored for expressing responsibilities and mandates [2].

B. Components of AI Systems

The ISO/IEC 22989:2022 standard provides a functional view of AI systems with input, model processing and outputs as the main concepts. For systems based on machine learning, the concepts of training data, machine learning and continuous learning are added [1]. AIA defines the various kinds of data used in AI systems and discusses models and types of systems.

3. INCLUDING AI SYSTEMS IN RISK MANAGEMENT

A. Risk Management Frameworks and Standards

Risk assessment guidelines are defined by the ISO 31000 risk management standard [5] and the National Institute of Standards and Technology (NIST) risk management framework (RMF) described in NIST SP 800-37 [6]. These are refined for information security by ISO/IEC 27005 [7] and for cybersecurity by the NIST cybersecurity framework (CSF) [8]. Furthermore, ISO/IEC 23984 [9] provides AI-specific guidance for risk management, as does the NIST AI RMF [10].

By design, these standards and frameworks accommodate a wide range of possible systems, making them too complex to deploy in small organizations, especially if the team has no previous experience in risk management.

Our methodology starts with a generic three-step risk management process: context establishment, risk assessment and risk mitigation. The scope is defined as IT systems extended to be AI systems, and it provides guidance on identifying the most critical AI risks requiring action. The methodology aligns with ISO 31000 and ISO/IEC 27005, but should also be adaptable to the NIST RMF and CSF. Thus, organizations that later plan to adopt a standard risk management system can incorporate the work done using our methodology.

B. AI Considerations During Context Establishment

During context establishment, interested parties (including hidden ones) and relevant assets are documented, the risk appetite of an organization is defined, and risk owners are determined. The internal, national and regulatory requirements of interested parties are identified. Risk consequence, likelihood and acceptance criteria are determined, and a risk management approach is chosen. In this step, the party building an AI system needs to identify:

1) the data subjects or data owners on whose data the models have been trained;
2) the party who trained the model;
3) the party who is running the service; and
4) the service user.

The legal rights, obligations and motives of all parties must be taken into account. Each stakeholder can bring new applicable standards and regulations that need to be considered. Stakeholders can be considered as part of the organization or not.

The organization needs to identify where different types of data (models and training, input and output data) and software components (training, inference and data ingestion systems) originate and what the data flow is among the different components. Some risks may result from engaging with certain kinds of data or systems, so this mapping is a prerequisite.

Table I shows a simple way to map the relations between stakeholders and AI system components. Such visibility tables give a visual overview of the components to which each stakeholder has access. In this simple example, we have three stakeholders: the end user, the service provider (providing the AI front-end), and the AI application programming interface (API) provider (training and providing the model). All the stakeholders can see the end users' input data and the model output. The service provider and the AI API provider have access to the service provider's business data. Only the AI API provider can see the details of the model.

TABLE I: EXAMPLE VISIBILITY TABLE

	User input	Service provider data	AI model	Output
End user	X			X
Service provider	X	X		X
AI API provider	X	X	X	X

C. AI Considerations During Risk Assessment

Risk is often expressed in terms of the likelihood of a threat materializing and the severity of the consequences. The risk assessment phase roughly consists of risk identification, risk analysis and risk evaluation. During risk identification, the risks are found, recognized and described. The risk owners are also defined for each identified risk. During analysis, the causes, sources and likelihood of each risk, and the likelihood and severity of the consequences are determined. During evaluation, the results of the analysis steps are compared with risk criteria, prioritized and considered for risk treatment.

AI risk assessment builds on the context identified in the previous step. For each AI system component, we assess the risk in the context of the related stakeholders. The identification of this relationship is straightforward based on the visibility table compiled in the context establishment phase. For each identified stakeholder–component pair, we consider risks from three categories – cybersecurity, regulatory and AI-specific risks. Cybersecurity risks focus on the confidentiality, integrity and availability of AI system components (software, data and services). Regulatory risks deal with legal obligations that apply to stakeholders operating AI systems (for AI-specific regulations) or their components (e.g., regulations on personal data, copyrighted data or critical infrastructure). Finally, we define AI-specific risks as risks connected to the specificity of the algorithms, the impact of AI systems on our society and the ethical aspects of deploying AI.

Table II gives examples of defining a risk through vulnerabilities and threats. For each threat, the organization must determine the likelihood of it materializing and the severity of the consequences for its environment. The likelihood and severity of the same event can vary for different organizations.

In addition, it can be helpful to compare the risks of different deployments to decide on a solution for an organization. For instance, while a cloud provider may offer a wider range of security controls than a small organization can deploy by itself,

making an organization dependent on the cloud creates availability risks, should the connection to the cloud be lost.

TABLE II: EXAMPLE VULNERABILITY AND THREAT TABLE

Object	Risk category	Vulnerability	Threat
Output	AI risk	Biased or damaged model	End user will get an output that will direct them to act in a damaging way
Input	Regulatory risk	Insufficient legal basis for personal data processing	Service provider faces legal action over infringement of data protection regulations
Language model	Cybersecurity risk	Faulty identity management	AI API provider loses access to their infrastructure, stopping inference services

D. AI Considerations During Risk Treatment

There are several ways of treating risks, such as risk avoidance, risk modification, risk retention and risk sharing. The method is chosen based on the outcomes of the risk assessment process. Based on the prioritized list of risks sent for treatment, a set of necessary cybersecurity, AI security and regulatory controls will be determined so that the results will meet the organization's risk acceptance criteria.

The risk treatment of AI systems does not have any special steps. The controls for AI-specific risks can be different from regular cybersecurity controls, but the treatment is generally still done in the same way.

4. CYBERSECURITY AND REGULATORY RISKS AND CONTROLS

A. Cybersecurity Risks and Controls

As AI systems are cyberphysical systems, most standard cybersecurity controls apply. All stakeholders and components of AI systems can be considered stakeholders and assets in IT systems. A catalogue of cybersecurity controls can be found, for instance, in ISO/IEC 27002 [11] and NIST SP 800-53 [12].

B. Regulatory Risks and Controls

The legal landscape related to AI systems is developing rapidly. Regulations are being developed in the United States [13] and China [14]. In 2023, the European Union (EU) reached an agreement on the structure for a legal framework for AI. The regulation builds on a risk-based approach and distinguishes four types of AI systems: prohibited

AI (systems that manipulate user vulnerabilities, e.g., social scoring AI systems), high-risk AI, limited-risk AI, and minimal-risk AI [2], [15]. Specific transparency and disclosure requirements are provided for general-purpose AI systems [15]. Exceptions apply to research and development, open-source, national security, and military use [2]. A proposal for a directive on adapting non-contractual civil liability rules to AI is awaiting agreement [16].

In addition to legislation specifically regulating AI systems, there are also norms concerning product safety [17], [18], data protection [19]–[22], intellectual property [23], [24] and cybersecurity [25], [26], that must be followed. Sector-specific norms (e.g., in financial services or healthcare) and legal requirements in individual states will also apply. Finally, ethical principles [27] have to be followed. Together with the agreements between parties, they form the legal framework in which the AI system operates.[1]

In the case of legal aspects, especially in terms of liability, the role of the person in the AI system must also be taken into account. For example, duties and responsibilities to ensure compliance with the EU AI regulation vary by their role [2]. When processing personal data, data protection roles must also be taken into account [19], [28].

Non-compliance with legal requirements may lead to sanctions, including fines or suspension of operations until deficiencies are eliminated, as well as reputational damage and a decrease in system users. Litigation may also ensue if the rights of individuals are violated.

Control measures include the following:

1) Before starting the activity, prepare a legal scoping report to understand the legal framework in which you are operating. Map all applicable legislation, agreements and terms of service provisions, and keep the document up to date.
2) Some AI systems may need *ex-ante* conformity assessments and risk assessments [2]. A data protection impact assessment should be completed before processing personal data in an AI system (GDPR Art. 35; Directive (EU) 2016/680 Art. 27).
3) Ensure that you have relevant agreements, consents and licences for processing data (whether personal, copyrighted or other) throughout the life cycle of the AI system.
4) Implement organizational and technical measures to ensure both physical and digital security of the AI system and data throughout the entire AI system life cycle. Use appropriate privacy-enhancing technologies [29], [30].

1 It is crucial to familiarize oneself with all legal and contractual requirements to ensure the legality of all activities. The provided list of legislation is not exhaustive, so the relevant applicable legal framework must be assessed in each specific case.

5) Understand how the AI system works (human oversight), ensure system reliability and accuracy, apply a risk management system and best data governance practices through the system life cycle, prepare technical documentation and keep it up to date, and provide appropriate instructions and explanations to system users.

5. AI-SPECIFIC RISKS AND CONTROLS

A. Attacks Against AI Systems

AI systems make decisions based on data. These decisions can be critically important, can be based on sensitive data (e.g., in healthcare), or might have to be made in a split second and therefore lack human oversight (e.g., in self-driving cars or drones). These peculiarities of AI systems imply that the additional consideration of AI-specific threats is necessary for a complete risk analysis. The following is based on the German Federal Office for Information Security's 'AI security concerns in a nutshell' [31] as well as the Open Worldwide Application Security Project (OWASP) Foundation's 'OWASP Top 10 for LLM Applications' [32].

Evasion attacks are attacks where the attacker attempts to manipulate the model to return unexpected, incorrect or malicious outputs. For example, prompt injection is an evasion attack against a large language model (LLM), in which the attacker attempts to obtain an unauthorized output or have the model perform unauthorized actions by carefully constructing a prompt [33]. This prompt could be constructed with natural language, or it could utilize techniques such as adversarial suffixing [34]. Similarly, image classification models may be vulnerable to adversarial examples, where the model is manipulated into predicting an incorrect class through slight perturbations of the input [35]. Vulnerabilities related to evasion attacks can carry over in the case of transfer learning [36].

Information extraction attacks are attacks in which the attacker attempts to learn or reconstruct sensitive information such as model weights or training data. In an attribute inference attack, the attacker attempts to infer a sensitive attribute about an identity present in the training data by comparing the statistical relationships between features observed in model outputs with those that have been observed in the real world. Similarly, in a membership inference attack, the attacker tries to gain information about an identity's presence in the training data [37], [38]. In the case of model theft, the attacker attempts to construct a shadow model by feeding it training data gathered from model outputs. Another type of information extraction attack is model inversion, where the attacker attempts to reconstruct elements from the model's training data based on its outputs [39], [40].

Poisoning and backdoor attacks are aimed at training data. By altering the training data of an image recognition or a text-to-image model to associate certain inputs with a particular or a random incorrect label, the model could be steered to misclassify when detecting a particular object or its performance could be degraded in general [41], [42]. A backdoor attack is a more sophisticated form of data poisoning, where the labels are manipulated only when a particular trigger is present in an input.

Denial-of-service attacks, which are a type of availability attack, have some peculiarities in the context of AI applications. In autoregressive LLMs, for example, the processing power required to respond to a query depends on the length and content of the query. Lacking input validation or output length limits, an LLM could be made to generate very long outputs using its maximum context length, using up computational and memory resources and degrading performance for other users.

B. AI-Specific Risks

The adoption of AI for social, governance and industrial purposes as well as increasing reliance on AI have caused the emergence of previously unforeseen risks and ethical challenges. These risks can be broadly grouped into algorithmic and societal risks. Algorithmic risks are risks that arise from the technical aspects of an AI system and its application. For example, the output of a model might be biased, inaccurate or harmful, resulting from mistakes or attacks during model training. Societal risks emerge from the wider effects of AI on society and the unpredictability of future developments. The ethical challenges of AI adoption are broadly related to questions such as the choice of value models (e.g., utilitarian calculus in self-driving cars) as well as possible negative externalities of AI use.

Examples of AI-specific risks include:

1) Algorithmic risks: An AI system might fail to generalize on real-world data, underperform and give bad outputs. This is a particularly serious risk in critical applications such as healthcare, and is in turn amplified if the model is not explainable or lacks human oversight. In addition, a system might give outputs that are harmful or dangerous and, given bias in training data, discriminatory.
2) Societal risks: AI systems can expand the scope of human agency. This empowers users to not only do a lot of good but also to potentially cause considerable harm. As the speed of AI adoption and development outpaces regulatory efforts, there are significant risks of misuse. For example, malicious actors could use AI to aid them in developing weaponry. AI can be used to generate believable, high-quality disinformation, undermining

trust in online media. Autonomous AI agents could develop into an artificial superintelligence, which could pose an existential threat to humanity.

3) Ethical challenges: The use and possible misuse of AI raises a number of ethical questions. For example, a self-driving car might have to decide whether it is more appropriate to put its driver or a pedestrian at risk. Another example is exploitative or addictive applications, because empowering them with AI could increase the harm they pose to mentally and socially vulnerable individuals. There are also concerns such as ownership of AI-generated content, the implications of using AI in the justice system, and the ethics of job loss and other socioeconomic risks caused by AI versus the opportunity cost of inhibiting its adoption.

C. AI-Specific Controls

Evasion attacks can be mitigated by validating input prompts and outgoing requests, and monitoring model responses. Adversarial suffixing in LLMs as well as adversarial examples in image models can be mitigated by not releasing model weights publicly. If the model is connected to data sources or applications, it should never have more permissions than the user querying it and the model should be considered an untrusted user. To mitigate indirect prompt injections, in which LLMs accept compromised input from an external source, output received from external sources should be monitored and validated.

Information extraction attacks can be avoided to some degree by ensuring that training data does not contain sensitive personal data. In addition, an LLM application should never expose sensitive information in the pre-prompt. The model does not fundamentally distinguish between the prompt and the pre-prompt, so it should always be assumed that the user is capable of extracting it.

Data poisoning and backdoor attacks can be mitigated by scrutinizing the training data, applying quality criteria to filter it and validating its supply chain [43]. In addition, data poisoning can be detected at inference time by testing model performance for specific input categories.

To mitigate denial-of-service attacks, inputs should be validated, resource usage and API rate per user should be limited, and resource use should be monitored.

To ensure that an AI model performs consistently, performance should be monitored over time and across a diverse set of input categories. Similarly, to mitigate algorithmic risks related to bias or harmful outputs, safety metrics should be included in the monitoring process. Training data should be as diverse as possible. In addition,

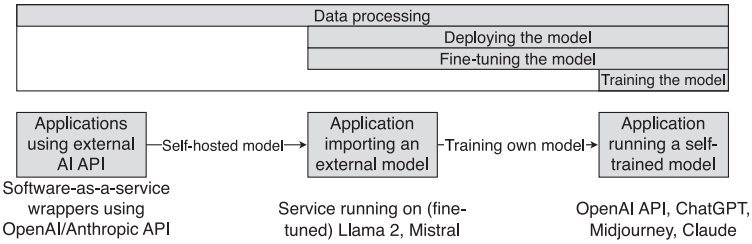
an effort should be made to make AI systems explainable, as explainability aids in interpreting monitored data, debugging the model and, thereby, achieving compliance.

6. GENERALIZED AI DEPLOYMENT BLUEPRINTS

A. Methodology

Architectural choices made in the design of an AI system significantly affect its risks. To simplify the risk analysis of new AI systems, we propose three generalized deployment blueprints (Figure 1). These blueprints differ in terms of the origin of the machine learning model, the party using the model on behalf of the service provider and data movement between parties. We consider AI systems with cloud services, as this is a common choice for AI systems where high computational performance is needed. However, the principles outlined here broadly apply to non-cloud applications as well. While these three deployment models do not cover every possible way to deploy an AI application, they can serve as guidance, helping application developers make sense of security and compliance risks.

FIGURE 1: OVERVIEW OF AI APPLICATION DEPLOYMENT MODELS WITH EXAMPLES



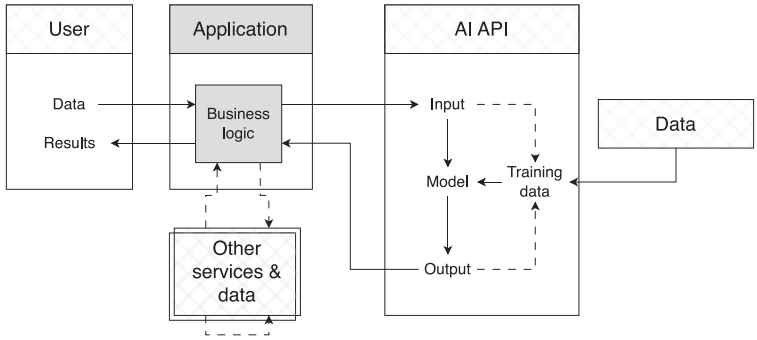
B. Systems Interfacing with an AI API

One way to deploy an AI system is by relying on a third-party API or AI-as-a-service in the business logic of the application (Figure 2). This permits the developer to leverage AI capabilities without having to deploy their own AI model for inference or having to train one on their own. In this deployment model, the system might also be interfacing with external data sources and services. The AI API might store the data it receives from the service to train its own models. This depends on the terms of service.

In this scenario, the AI model is external (not provided by the service provider), as is the training data. If the applications process user data, then user data moves to the application service, and from there to the AI API. The output is returned to the application service and then passed to the user after intermediate processing. It is

possible that user data is stored by both the service provider and the AI API provider. It is important to consider that relying on an external API externalizes availability risks. In addition, not all AI API providers support adapting (e.g., fine-tuning) the model to the service providers' needs.

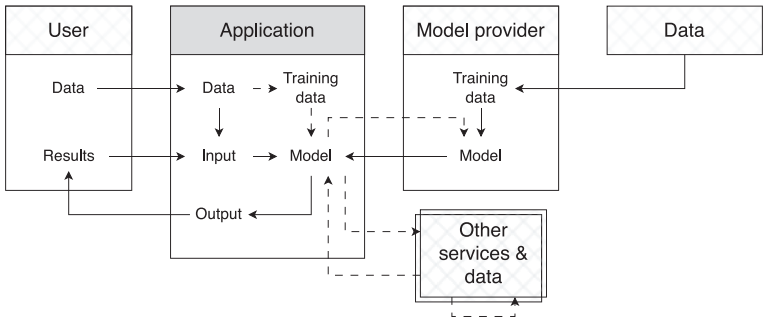
FIGURE 2: AI SYSTEMS RELYING ON AN EXTERNAL AI API



C. Systems Using a Self-Hosted External AI Model

By self-hosting a pre-trained model obtained from a model provider and (if necessary) adapting and fine-tuning it, some control over the deployment is regained (Figure 3). However, this deployment model introduces new challenges: the AI system owner now has to procure appropriate hardware and be more responsible for the model outputs, security and safety properties. In the case of fine-tuning, the AI system provider has to manage and curate the training data. In this scenario, user data is only stored by the service provider, unless third-party processors (e.g., cloud) are integrated into the system.

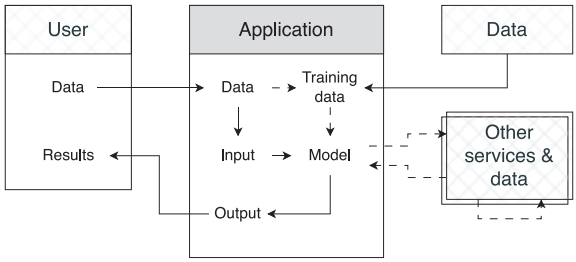
FIGURE 3: AI SYSTEMS IMPORTING AND SELF-HOSTING AN EXTERNAL MODEL



D. Systems Training Their Own Models

If an AI system is deployed using only self-trained and self-hosted AI models, risks related to data processing are internalized, as are those related to model performance and data quality (Figure 4). This deployment model is also used by dedicated AI technology providers who have the resources to deploy everything in-house, as well as by parties deploying the simplest, least computation-intensive AI applications, which do not require specialized hardware to train or host. Here, most data can be processed by the service provider.

FIGURE 4: AI SYSTEMS SELF-HOSTING A SELF-TRAINED MODEL



7. RISKS FOR AI SYSTEMS BASED ON THE BLUEPRINTS

Each of the presented three blueprints considers the system from the service provider's standpoint. Thus, the scope of the AI system under the control of the service provider increases. Service providers using an AI API are indirectly affected by risks related to inference or training. However, choices made by the stakeholders operating the inference and training components affect the service provider.

We provide a list of key risks to consider for service providers operating each blueprint. Once a service provider has identified that it is building an AI system that matches one of the three blueprints described in the previous section, it can make use of Table III to identify the key risks affecting that design.

The risks are grouped according to the three common stages in an AI workflow. The service making use of AI needs to process input data and output data using certain business logic on some infrastructure. This stage is shared by all blueprints. If the service provider opts to run inference with the AI model itself in order to limit data transfers to third parties, certain risks may be reduced, but the model needs to be

selected carefully. Finally, if the service provider trains the model itself, it needs to consider the risks related to processing training data and model quality. In the other blueprints, the service provider can try to make the AI API or AI model provider contractually liable for these concerns.

It must be noted that the risks listed in the table are the ones specific to AI systems. In each blueprint, common cybersecurity risks need to be assessed for each component for which the service provider is responsible. Also, each territory may have applicable local legislation or standards regulating the use of AI systems. These may add additional risks not included in the table.

We foresee that a user of this methodology will benefit from supportive tools (forms, tables, figures and worksheets) that help organize the information needed to follow the methodology. We have designed the first versions of such forms and the following guidance. These have not been included in this paper due to their size and are available in a separate report [44].

After performing an initial risk assessment using this table and picking the relevant mitigations, the service provider can set out to perform a full risk assessment according to a framework of their choice. The risk management conducted according to the blueprints in this paper will contribute to that analysis and ensure that it starts from a strong basis.

TABLE III: KEY RISKS FOR THE THREE GENERALIZED BLUEPRINTS

		External AI API	Self-hosted AI model	Self-trained AI model
Service provision	**Cybersecurity risks**	Availability of the AI API is not under the control of the service provider.	*Standard risks apply.*	*Standard risks apply.*
	Regulatory risk	Service provider does not have rights to process data, e.g., for transfer to the AI API provider, cloud or across borders.	Service provider does not have rights to process data, e.g., for transfer to cloud or across borders.	Service provider does not have rights to process data, e.g., for transfer to cloud or across borders.
	AI-specific risks	AI API uses a model that produces outputs that are unsafe or leak data.	*See inference risks below.*	*See inference risks below.*
Inference	**Cybersecurity risks**	*Not in scope of the service provider.*	Infrastructure used for AI inference does not perform well enough. Model provider does not provide updates.	Infrastructure used for inference does not perform well enough.
	Regulatory risks	*Not in scope of the service provider.*	Model contains data for which service provider does not have processing rights. Service provider does not have rights to process fine-tuning data.	*See training risks below.*
	AI-specific risks	*Not in scope of the service provider.*	AI model produces outputs that are unsafe or leak data. Data and tools used for fine-tuning reduce model quality.	*See training risks below.*
Training	**Cybersecurity risks**	*Not in scope of the service provider.*	*Not in scope of the service provider.*	Infrastructure used for model training does not perform well enough.
	Regulatory risks	*Not in scope of the service provider.*	*Not in scope of the service provider.*	Service provider does not have rights to process training data.
	AI-specific risks	*Not in scope of the service provider.*	*Not in scope of the service provider.*	AI model produces outputs that are unsafe or leak data. Data and tools used for fine-tuning reduce model quality.

8. FUTURE WORK

Further research includes comparing the proposed methodology with other lightweight risk management methodologies to quantitatively measure the effort needed for initial application and later maturation to a full quality management system. This will include a qualitative evaluation combining analysis of interview results with the evaluation of self-made assessments by professional risk managers.

A full report that details the background material and proposed methodology and provides supporting worksheets and a user's guide has been published by the Estonian Information System Authority [44]. Thus, we expect the methodology to find real-world use in both the public and private sector, providing opportunities to continue the suggested research.

REFERENCES

[1] *Artificial Intelligence – Artificial Intelligence Concepts and Terminology*. Standard ISO/IEC 22989:2022, ISO, 2022.

[2] 'Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act)', 2021/0106(COD), European Commission, 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206

[3] 'Recommendation of the Council on artificial intelligence. Amended on: 08/11/2023', *OECD Legal Instruments*, OECD, 2023.

[4] White House Office of Science and Technology Policy, 'Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People', The White House, 2023.

[5] *Risk Management – Guidelines*. Standard ISO 31000:2018, ISO, 2018.

[6] *Risk Management Framework for Information Systems and Organizations: A System Life Cycle Approach for Security and Privacy*. Standard NIST SP 800-37 Rev. 2, US NIST, 2018.

[7] *Information Security, Cybersecurity and Privacy Protection – Guidance on Managing Information Security Risks*. Standard ISO/IEC 27005:2022, ISO, 2022.

[8] NIST Cybersecurity Framework 1.1. NIST, 2018.

[9] *Artificial Intelligence – Guidance on Risk Management*. Standard ISO/IEC 23984:2023, ISO, 2023.

[10] NIST AI Risk Management Framework 1.0. NIST, 2023.

[11] *Information Security, Cybersecurity and Privacy Protection – Information Security Controls*. Standard ISO/IEC 27002:2022, ISO, 2022.

[12] *Security and Privacy Controls for Information Systems and Organizations*. Standard NIST SP 800-53 Rev. 5. NIST, 2020.

[13] 'Fact sheet: President Biden issues executive order on safe, secure, and trustworthy artificial intelligence'. The White House. 30 Oct. 2023. [Online]. Available: https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artif

[14] 'Interim Measures for the Management of Generative Artificial Intelligence (AI) Services'. Cyberspace Administration of China. 13 Jul. 2023. [Online]. Available: http://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm

[15] 'AI Act: Deal on comprehensive rules for trustworthy AI'. European Parliament. 9 Dec. 2023. [Online]. Available: https://www.europarl.europa.eu/news/en/press-room/20231206IPR15699/artificial-intelligence-act-deal-on-comprehensive-rules-for-trustworthy-ai

[16] 'Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)', 2022/0303(COD), European Commission, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496

[17] Regulation of the European Parliament and of the Council of 10 May 2023 on General Product Safety, *Official Journal*, L 135, 23 May 2023, pp. 1–51. [Online]. Available: http://data.europa.eu/eli/reg/2023/988/oj
[18] 'Proposal for a Directive of the European Parliament and of the Council on liability for defective products', 2022/0302(COD), European Commission, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52022PC0495
[19] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016, (General Data Protection Regulation), *Official Journal*, L 119, 4 May 2016, pp. 1–88. [Online]. Available: https://eur-lex.europa.eu/eli/reg/2016/679/oj
[20] 'Proposal for a Regulation of the European Parliament and of the Council laying down additional procedural rules relating to the enforcement of Regulation (EU) 2016/679)', 2023/0202(COD), European Commission, 2023. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52023PC0348
[21] Directive (EU) 2016/680 of the European Parliament and of the Council of 27 April 2016, *Official Journal*, L 119, 4 May 2016, pp. 89–131. [Online]. Available: http://data.europa.eu/eli/dir/2016/680/oj
[22] Regulation (EU) 2018/1725 of the European Parliament and of the Council of 23 October 2018, *Official Journal*, L 295, 21 Nov. 2018, pp. 39–98. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32018R1725
[23] Berne Convention for the Protection of Literary and Artistic Works (as amended on September 28, 1979). [Online]. Available: https://www.wipo.int/wipolex/en/text/283693
[24] WIPO Copyright Treaty. [Online]. Available: https://www.wipo.int/wipolex/en/text/295157
[25] Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022, *Official Journal,* L 333, 27 Dec. 2022, pp. 80–152. [Online]. Available: http://data.europa.eu/eli/dir/2022/2555/oj
[26] 'Proposal for a Regulation of the European Parliament and of the Council on horizontal cybersecurity requirements for products with digital elements and amending Regulation (EU) 2019/1020', European Commission, 2022. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022PC0454
[27] European Parliament Resolution of 20 October 2020 with Recommendations to the Commission on a Framework of Ethical Aspects of Artificial Intelligence, Robotics and Related Technologies (2020/2012(INL)), *Official Journal*, C 404, 6 October 2021, pp. 63–106. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0275
[28] 'Guidelines 07/2020 on the concepts of controller and processor in the GDPR', European Data Protection Board, 7 Jul. 2021. [Online]. Available: https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-072020-concepts-controller-and-processor-gdpr_en
[29] D. Bogdanov, E. Brito, P. Etti, L. Kamm, P. Laud, T. Mällo, A. Ostrak, K. Sein, R. Talviste, and M. Toomsalu. 'Privacy enhancing technology concept', (in Estonian), Cybernetica, Estonian Ministry of Economic Affairs and Communications, Tallinn, Estonia, 31 Mar. 2023. [Online]. Available: https://www.kratid.ee/_files/ugd/980182_f1288bebbb57466ead0241748d49d8ec.pdf
[30] D. Bogdanov, E. Brito, P. Etti, L. Kamm, P. Laud, T. Mällo, A. Ostrak, K. Sein, R. Talviste, and M. Toomsalu. 'Roadmap for deploying privacy enhancing technologies in Estonia', (in Estonian), Cybernetica, Estonian Ministry of Economic Affairs and Communications, Tallinn, Estonia, 31 Mar. 2023. [Online]. Available: https://www.kratid.ee/_files/ugd/980182_64478f7163b74f299f5879b6eea856af.pdf
[31] 'AI security concerns in a nutshell – practical AI-security guide', Federal Office for Information Security, Bonn, Germany, 2023.
[32] 'OWASP top 10 for large language model applications. Version 1.1', OWASP Foundation, 2023.
[33] Y. Liu, G. Deng, Y. Li, K. Wang, T. Zhang, Y. Liu, H. Wang, Y. Zheng, and Y. Liu. 'Prompt injection attack against LLM-integrated applications', 2024, *arXiv.2306.05499*.
[34] A. Zou, Z. Wang, N. Carlini, M. Nasr, J. Zico Kolter, and M. Fredrikson, 'Universal and transferable adversarial attacks on aligned language models', 2023, *arXiv.2307.15043*.
[35] I. J. Goodfellow, J. Shlens, and C. Szegedy, 'Explaining and harnessing adversarial examples', 2014, *arXiv.1412.6572*.
[36] J. Lin, L. Dang, M. Rahouti, and K. Xiong, 'ML attack models: Adversarial Attacks and Data poisoning attacks', 2021, *arXiv.2112.02797*.
[37] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, 'Membership inference attacks against machine learning models', *IEEE Symposium on Security and Privacy (SP)*, USA, pp. 3–18, 2017, doi: 10.1109/SP.2017.41.
[38] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar, 'Membership inference attacks against synthetic data through overfitting detection', 2023, *arXiv.2302.12580*.

[39] N.-B. Nguyen, K. Chandrasegaran, M. Abdollahzadeh, and N.-M. Cheung, 'Re-thinking model inversion attacks against deep neural networks', *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Canada, pp. 16384–16393, 2023, doi: 10.1109/CVPR52729.2023.01572.
[40] K.-C. Wang, Y. Fu, K. Li, A. Khisti, R. Zemel, and A. Makhzani, 'Variational model inversion attacks', 2022, *arXiv.2201.10787*.
[41] S. Shan, W. Ding, J. Passananti, H. Zheng, and B. Y. Zhao, 'Prompt-specific poisoning attacks on text-to-image generative models', 2024, *arXiv.2310.13828*.
[42] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, and T. Goldstein. 'Dataset security for machine learning: Data poisoning, Backdoor Attacks, and Defenses', *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 45, no. 2, pp. 1563–1580, 2023, doi: 10.1109/TPAMI.2022.3162397.
[43] I. Brown. 'Expert explainer: Allocating accountability in AI supply chains'. 2023. [Online]. Available: https://www.adalovelaceinstitute.org/resource/ai-supply-chains
[44] D. Bogdanov, P. Etti, L. Kamm, A. Ostrak, F. Stomakhin, M. Toomsalu, S.-M. Valdma, and A. Veldre 'A study of the risks and controls for artificial intelligence and machine learning technologies 1.0', (in Estonian), Cybernetica, Estonian Information System Authority, 27 Feb 2024. [Online]. Available: https://www.ria.ee/sites/default/files/documents/2024-03/Tehisintellekti-masinoppe-tehnoloogia-riskide-uuring-2024.pdf