Cyber Risk Challenges for AI: An Explorative Study into Cyber Risk Management for AI-powered Systems in the Dutch Financial Sector

Mol, Yuri

A thesis submitted in fulfilment of the requirements of the Master of Science: Executive Master's Programme in Cyber Security degree, Faculty of Governance and Global Affairs.

Cyber Risk Challenges for AI: An Explorative Study into Cyber Risk Management for AI-powered Systems in the Dutch Financial Sector

---

17.825 words

Author:
Yuri Alexander Mol

Supervisor:
Dr. Els de Busser

Second Reader:
Dr. Bibi van den Berg

LEIDEN UNIVERSITY

Amsterdam, the Netherlands
6 January 2023

Abstract

Artificial intelligence (AI) and cyber security failure are two of the highest impact risk areas of this decade, with developments surrounding these two topics going hand in hand. For instance, AI can act as force multiplier for existing cyber threats and can be an enabler for new cyber threats. Moreover, AI can bolster cyber defences, and AI-powered systems can provide a new attack surface for cyber threats. Due to increased investment into and utilisation of AI, rising cyber threats, a growing cyber risk awareness in society and resultant regulatory, governance and technical risk treatment efforts, the last area of concern is especially relevant. In this regard, the financial sector is of specific interest due to its high investments into AI and cyber risk management, as well as high exposure to cyber risks. Financial institutions active in the EU also must consider upcoming legislation like the Digital Operational Resilience Act (DORA) and AI Act, which require them to adequately manage cyber risks for their AI systems.

In contrast to more established technologies, there is no universal cyber risk management or security control framework designed for AI-powered systems, and such systems pose unfamiliar risks and provide new challenges. This is problematic since appropriate guidance is important for the implementation of effective cyber risk management practices. Furthermore, without insight in current cyber risk management practices for AI-powered systems, it is difficult to determine whether legislation or guidelines are fit for purpose, which is of importance from a regulatory perspective. As such, this thesis sets out to study which cyber risks AI-powered systems face, how cyber risks for AI-powered systems are managed in the financial sector, and which internal governance and control practices are used in relevant cyber risk management processes. The research is primarily based on a literature review, with the supportive method being semi-structured interviews. For scoping purposes, the interviews were conducted with experts and practitioners active in the Dutch financial sector. Still, due to the cross-border nature of the sector, results are likely to be applicable to the broader European financial sector.

The five main findings are that: 1) While regulatory developments like DORA and the AI Act have resulted in increased financial sector attention for cyber risk management for AI-powered systems, the current state of play is that AI use is not widespread, with AI complexity being low and it often being used in a low-risk environment, resulting in relevant cyber risk management practices not being top of mind; 2) important focus areas in cyber risk management for AI systems are data and model risk; 3) next to AI data and model dependencies, AI system interconnectivity is another important source of risk, resulting in AI supply chain risks being an important focus area for cyber risk management; 4) concerning cyber risk identification and analysis practices, interview findings show that any cyber risk management for AI system framework should use an ecosystem perspective that considers the environment in which the system and organisation operate in, and 5) while increased regulatory attention to cyber risk management for AI is generally seen as a good thing, the multifaceted characteristics of AI systems and the risks they face require due consideration of potential contradictory regulatory requirements.

Table of contents

1. Introduction

1.1. Overview

Artificial intelligence (AI) and cyber security failure are identified as two of the highest impact risk areas of this decade, with developments surrounding AI and cyber security going hand in hand.[1] To ensure the safe development and use of AI, cyber security is an important precondition. This is also true vice versa, to defend against emerging cyber security threats, it is necessary to closely follow developments in AI.[2] In general, this relationship can be divided into five areas of concern, namely: 1) AI can act as force multiplier for existing cyber security threats, with AI transforming the range and reach of threats like malware; 2) AI is an enabler for emerging cyber security threats, with AI creating new threats, such as deepfakes and AI-fused data for phishing purposes; 3) AI-powered systems (hereinafter also referred to as AI systems) and their assets provide a new attack surface for cyber security threats, with AI being susceptible to, among others, training data manipulation and data lake poisoning; 4) AI can act as catalyst for cyber security threats, which includes specific scenarios like rapid machine-to-machine escalation via automated command-and-control servers, and 5) AI-enabled cyber defences provide new capabilities in, inter alia, flagging and blocking connections and communications missed by current intrusion detection and prevention technologies.[3]

Due to a combination of several factors, such as increasing investments into and utilisation of AI systems, rising cyber security threats and a related growing cyber risk awareness in society, resulting in heightened regulatory, governance, and technical risk treatment efforts, the third area of concern, AI-powered systems providing a new attack surface for cyber security threats, is worth examining in closer detail.[4] One of the industry sectors that is of specific interest in this regard is the financial sector, a sector that has seen expanding investments into AI, is increasingly targeted in cyberattacks, is strongly regulated, and in which risk management is a key focus area for doing business. For example, the Artificial Intelligence Index Report 2022 shows that next to high-tech/telecommunications, financial services show the greatest adoption by industry. [5] New technologies like AI increase financial sector reliance on IT, resulting in a growing exposure to cyber risks.[6] This is concerning because cyber-related operational losses in the financial sector are increasing, with Dutch banks even doubling losses between 2018 and 2020, and 5% of Dutch pension funds and insurers already being victims of a successful cyberattack in 2021.[7]

While growing operational losses and the implementation of appropriate risk management is challenging, another key challenge is that it is not clear what appropriate risk management should look like.[8] Cyber risk management, especially that of new technologies such as AI is highly complex. This contrasts with traditional risk management practices, which focus on conventional risks that are more easily isolated. Cyber risks are trans-boundary and do not have a singular root cause. This is further complicated by new technologies like AI that are characterised by their complexity and

---

[1] World Economic Forum, The Global Risks Report 2022: 17th Edition Insight Report, pp. 7, 47-53.
[2] WRR, Opgave AI: De nieuwe systeemtechnologie, 5 November 2021; CSR, Adviesrapport Integrale Aanpak Cyberweerbaarheid, 2021.
[3] National Security Commission on Artificial Intelligence (NSCAI), Final Report, 2021, pp. 45, 278-281.
[4] *See* World Economic Forum, *supra* note 1, at 27.
[5] The AI Index 2022 Annual Report, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022, p. 161.
[6] BIS, Working Papers No 865: The drivers of cyber risk (May 2020), pp. 2-10.
[7] DNB, a macroprudential perspective on cyber risk, Occasional Studies Volume 20 – 1 (2022), p. 4; (https://www.dnb.nl/nieuws-voor-de-sector/toezicht-2022/dnb-ziet-cyberdreiging-toenemen-terwijl-basismaatregelen-niet-altijd-op-orde-zijn/), last visited (8-12-2022); (https://www.bankingsupervision.europa.eu/banking/srep/2021/html/ssm.srep202107_outcomesrepitriskquestionnaire.en.html#toc4), last visited (8-12-2022).
[8] L. Mauri and E. Damiani, Modeling Threats to AI-ML Systems Using STRIDE, Sensors (2022), pp. 17-18; X. Zhang, F.T.S. Chan, C. Yan, and I. Bose, Towards risk-aware artificial intelligence and machine learning systems: An overview, Decision Support systems (2022), p. 9.

interconnectedness and are subject to a fast-changing threat landscape. As such, organisations exposed to cyber risks need to apply risk management practices that are both robust, comprehensive, and flexible, and that can deal with a complex technological and risk environment. Within this context, the financial sector warrants special attention because it is held to a higher societal standard, with trust being essential to the functioning of the financial systems. Moreover, the financial sector has systemic importance, resulting in disruptive technologies like AI causing potential systemic risks.[9] In addition, the financial sector has a unique data environment characterised by a prevalence of data being available, which makes it well suited for the implementation of AI systems.[10]

Financial sector cyber risk exposure and the importance of risk management of new technologies like AI systems has also been picked up by EU regulators, with the upcoming regulation on digital operational resilience for the financial sector (DORA) and for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act or AI Act), requiring financial institutions to have in place an internal governance and control framework that ensures effective management of all IT risks (e.g. Article 4(1) DORA) and that high-risk AI systems shall be designed and developed to achieve an appropriate level of accuracy, robustness and cyber security (e.g. Article 15(1) AI Act).[11] Regulatory changes, and AI and cyber security developments have caused the management of related risks to become a board-level concern. However, it is unclear to what extent Dutch financial institutions are aware of the specific cyber risks their AI systems face and how they manage these risks. As such, this thesis sets out to study which cyber risks AI-powered systems face, how cyber risks for AI-powered systems are managed in the Dutch financial sector, and which internal governance and control practices are used in relevant cyber risk management processes.

This thesis is primarily based on a literature review, with the secondary data source being semi-structured interviews. These interviews enable a cross-comparison of findings and advances made by the academic research community to the current state of play in the Dutch financial sector, providing insight into gaps and potential areas for improvement in future research and/or cyber risk management practices in the sector. For the purposes of this study, cyber risk management is defined as the risk management process, consisting of, inter alia, risk identification, risk analysis, and risk treatment steps, used to manage cyber risk.[12] Cyber risk is then defined as:

"… an operational risk associated with performance of activities in the cyberspace, threatening information assets, ICT resources and technological assets, which may cause material damage to tangible and intangible assets of an organisation, business interruption or reputational harm. The term 'cyber risk' also includes physical threats to the ICT resources within organisation."[13]

Last, AI-powered systems are interpreted in the broadest sense, meaning any machine or system trained to perform tasks independently that would ordinarily require biological brainpower to accomplish. In this regard, machine learning (ML), neural networks, deep learning, and other AI-related concepts are simply treated as subsets of AI and are therefore not treated separately.[14] For the financial sector,

---

[9] ESRB, Mitigating systemic cyber risk, January 2022, pp. 4-25.
[10] DNB, General principles for the use of Artificial Intelligence in the financial sector, 2019, pp. 30-31.
[11] Regulation of the European Parliament and of the Council on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014 (hereinafter referred to as DORA) (version: compromise text 23 June 2022); (https://www.consilium.europa.eu/en/press/press-releases/2022/11/28/digital-finance-council-adopts-digital-operational-resilience-act/), last visited (8-12-2022); Proposal for a 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 (hereinafter referred to as AI Act) (version: compromise text 15 July 2022); (https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/), last visited (8-12-2022).
[12] M. W. Elliott, Risk in an evolving world, The Institutes (2019), pp. 1.23-1.27; (https://www.iso.org/iso-31000-risk-management.html), last visited (20-12-2022).
[13] G. Strupczewski, Defining cyber risk, Safety Science 135 (2021), p. 6.
[14] (https://www.turing.ac.uk/about-us/frequently-asked-questions), last visited (28-10-2022).

this means that machine learning tools and systems used for, inter alia, customer interaction, transaction monitoring and detection, and anomaly detection fall within the scope of this definition.

1.2. Research objective and questions

The purpose of this study is to gain insight into the current state of play of cyber risk management for AI-powered systems in the Dutch financial sector and to explore the demands that new technologies such as AI place on cyber risk management practices. The primary research question of this thesis is:

> "Which cyber risks do AI-powered systems face and how does the Dutch financial sector manage these risks?"

In answering this question, the following sub-questions will be considered:

i. Which cyber risks do AI-powered systems face?
ii. How can cyber risks for AI-powered systems be managed?
iii. What risk management practices do Dutch financial institutions apply to manage cyber risks for AI-powered systems?

1.3. Theoretical framework

Socio-technical, economic, and regulatory developments increase the need for robust cyber risk management practices for AI-powered systems but in contrast to more established technologies, there is currently no universal cyber risk management or security control framework specifically designed for these systems. This is a problem because appropriate guidance is important for the implementation of effective cyber risk management practices. For instance, no threat identification method is practical without some form of established principles in selecting controls to mitigate identified threats. Therefore, there is a need for further research into how conventional cyber risk management practices can be complemented by AI-oriented risk management practices, such as exploring how widely used standards like those published by NIST or the ISO can be applied in a manner to effectively mitigate risks for AI-powered systems.[15]

The absence of a cyber risk management framework dedicated to AI systems is also problematic since these systems pose unfamiliar risks and provide challenges that are distinct from those before. Moreover, with reliance on AI systems continuing to grow and there being no well-established risk management models and frameworks, risks for AI systems will turn into systemic risks that can threaten the stability and well-functioning of the Dutch financial sector. As such, there is a need for a cyber risk management model and framework that accommodates the characteristics of AI-powered systems and their specific risks.[16] In addition, due to the absence of a universal holistic theoretical framework for AI cyber risk management, it is difficult to determine what is excessive or missing in existing and upcoming legislation and guidelines. This is an issue from a regulatory perspective because it is difficult to determine whether legislation or guidelines are fit for purpose.[17]

The results of this explorative study into cyber risks for AI-powered systems and relevant cyber risk management practices in the Dutch financial sector can be used to conceptualise what a relevant cyber risk management framework for AI systems should look like. Furthermore, the results of this study provide insights in what future research directions could be considered.

---

[15] *See* Mauri *et al*, *supra* note 8, at 17-18.
[16] X. Zhang, F.T.S. Chan, C. Yan, and I. Bose, Towards risk-aware artificial intelligence and machine learning systems: An overview, Decision Support systems (2022), p. 9.
[17] K. Jia and N. Zhang, Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines, Electronic Markets (2022), p. 69.

1.4. Methodology

This thesis is qualitative in nature, providing an explorative study based on a literature review and semi-structured interviews. A qualitative approach is chosen because: 1) there is a lack of empirical studies and hard data on cyber risks for AI-powered systems, and 2) (cyber) risk management is largely context/environment specific and is mostly based on professional judgement and educated guesses rather than uniform rules and verifiable data. Moreover, such an approach provides flexibility to adjust based on new insights gained during the term of the study, which is useful when dealing with a relatively novel phenomenon. To enhance the validity and credibility of the findings, the study applies data triangulation, which refers to using multiple data sources, in this case, primary academic and secondary sourced literature and interviews, to develop a comprehensive understanding of the phenomenon.[18] In addition, template analysis is used for coding the qualitative data.[19] The primary method is a literature review of recent academic research on the topic of cyber risk management for AI-powered systems and assets. The supportive method is semi-structured interviews conducted with cyber security and risk management experts active in the Dutch financial sector.

1.4.1. Literature review

The thesis is primarily based on a literature review of primary sources on traditional risk management practices and cyber risk management practices for new technologies in general and AI-powered systems in particular, as well as secondary sources, such as guidance from cyber security and AI related authorities, agencies, expert opinions and white papers. Literature was collected through the digital library of Leiden University as well as other academic search engines like Google Scholar, using key words, such as ‘cyber risk management’, ‘AI risk framework’, and ‘cyber risk management for AI’. Articles were then selected based on their relevance to the research scope and theme. The literature review provides a conceptualisation of cyber risk management for AI-powered systems, contributing towards a better understanding of the research topic and definition of a priori themes for the qualitative data coding. In preparation of the template analysis method, three a priori themes were identified: 1) technological characteristics; 2) cyber risks, and 3) cyber risk management. These a priori themes were selected based on their relevance to the research question and for structuring the initial coding phase of analysis.[20]

1.4.2. Interview approach

Interviews were conducted using a semi-structured approach, covering all questions in every interview, which were used for follow-up questions with the interviewees. The interviews were conducted through a video call, with an interview lasting between 30 minutes and one hour. At the start of the interview, participants were informed of its confidential nature and that results would be processed anonymously, as well as that questions were answered on an entirely voluntary basis. Where participants gave explicit permission, interviews were recorded, and full interview transcripts were made. In case participants did not provide permission, notes were taken, which were then reviewed by the participant. In total, out of five interviews, only the interview conducted on 10 November 2022 was not recorded, with the interviewee providing their notes. In preparation for the interviews, two types of (largely similar) interview templates were created, one for practitioners employed at a Dutch financial institution and one for subject experts, which were shared with the participants beforehand.

Two interview templates were chosen because practitioners were requested to primarily answer questions that provided insight into the cyber risk management process of their financial institution, while experts were mainly asked questions meant to provide insight into the more macro-level state of play of

[18] N. Carter, D. Bryant-Lukosius, A. DiCenso, J. Blythe, and A. J. Neville, The Use of Triangulation in Qualitative Research, Methods & Meanings, Oncology Nursing Forum (2014), p. 545.
[19] N. King and J.M. Brooks, Template Analysis for Business and Management Students, SAGE Research Methods 2017.
[20] Id, pp. 25-46.

the sector. The templates were shared beforehand because all interviewees wanted access to the interview questions as a condition for participation since prior permission was needed from their legal/compliance department and/or their management. The interview questions were divided into three categories: background questions, macro-level questions, and micro-level questions. For interviewees employed at Dutch financial institutions, the background questions focused on gaining information on the background, function, role, and position in the financial institution of the interviewee. For expert interviewees (e.g. academic researchers, consultants or regulators), the background questions focused on the function, specific area of expertise and previous relevant work experience of the interviewee.

For practitioners, macro-level questions were directed at gaining insight into what kind of AI systems the financial institution used, what the AI asset taxonomy looked like, and how relevant vulnerabilities and threats targeting these vulnerabilities were identified, managed, and prioritised. Macro-questions were also used to gain insight into (executive) management support and other relevant governance or technical cyber risk management developments. For subject experts, macro-level questions were directed at gaining a better understanding of what kind of AI systems are in use in the Dutch financial sector, what important AI assets can be found, and how relevant vulnerabilities and threats targeting these vulnerabilities can be identified, managed, and prioritised. For both types of interviewees, micro-level questions aimed to gain more understanding of the cyber security and AI life cycle management decision-making process, information sources and metrics used, and the implementation of cyber risk management frameworks or industry best practices.

1.4.3. Interview population

Although more than 20 potential interviewees were approached in the period September-December 2022, at the end only five (3) practitioners and (2) experts were available. The three interviewed practitioners were employed at Dutch financial institutions that play an important role in the country's financial sector due to their size, complexity and/or overall risk profile; the two experts were employed in relevant regulatory and/or supervisory authorities. The interviewees all had a background in either cyber security, AI, data science, (IT) risk management and audit, or a combination thereof, and were all actively involved in cyber risk management for AI-powered systems.

1.4.4. Data analysis

The qualitative data is coded and analysed through the template analysis method, which is a research approach where qualitative data is systematically analysed through hierarchical coding via a coding template to identify key themes. Frequency and pattern of theme distribution then highlight areas of potential interest for further research.[21] For this study the following steps were undertaken:[22]

1. Familiarisation with data through thorough reading of interview transcripts and notes.
2. Preliminary or first cycle coding, using descriptive coding in which important words and phrases are identified and indexed using the a priori themes.
3. Clustering of the indexed data and second cycle coding, grouping the indexed data using theoretical coding in which shared keywords or key phrases are identified that trigger a discussion of the research question.
4. Drafting an initial linear style template, in which the second cycle coding outcomes were used to formulate, categorise, and rank themes based on their prevalence and relevance to the research question.
5. Applying and developing the template, in which the qualitative data is linked to the top themes, providing a coding hierarchy.
6. Final interpretation, in which the top-level themes and coding hierarchy are used to create an integrative narrative.

[21] Ibid.
[22] Ibid.; J. Saldaña, The Coding Manual for Qualitative Researchers, SAGE 2016, pp. 102-104, 250-254.

Last, the findings in the literature review and the analysis of the qualitative interview data are used to examine the current state of play regarding cyber risk management of AI-powered systems in the academic research community vis-à-vis practice in the Dutch financial sector.

1.4.5. Strengths and limitations

Regarding the interview method, the advantage of the semi-structured approach is that it enables the gathering of contextual information, providing the opportunity to gain new or unexpected insights; something that would not be possible using a structured interview approach. The disadvantage of the chosen methodology is that the findings are not generalisable, which means that they are explorative in nature. However, since the use of AI systems and the need for managing relevant cyber risks in the Dutch financial sector is a novel phenomenon, the semi-structured interview methodology is justifiable. In further research, the findings in this thesis can be tested using a structured interview approach or other more quantitative methods to gain a deeper understanding of the issues at play. In addition, another limitation of the interview findings is the small sample size (5) and questions having to be shared beforehand, resulting in interviewees potentially being primed. These limitations were partially overcome by interviewees holding senior positions within organisations that play an important role in the Dutch financial sector, meaning that despite the small sample size, interview findings likely reflect the broader state of play in the sector, and by the open-ended nature of most interview questions.

Concerning the template analysis method for data coding, its strengths include adaptability, flexibility, efficiency, and transparency. This allows for a certain amount of freedom in how the core features of template analysis are applied, providing room to style the research method in a manner that best fits the research subject and environment. Moreover, while the development of the initial template is done thoroughly, subsequent coding and further improvement of the template require less time than other related methods. In addition, depth of analysis is not prescribed, meaning that application can be fitted to the resources and time available, and the coding template itself provides an audit trail of the analytical process.[23]

Limitations and challenges of template analysis include its generic nature, fragmenting accounts, and limited guidance on final interpretation. It is generic in the sense that it is not tied to a specific philosophical or theoretical position. This is left to the researcher, which means that analytical decisions and interpretation of data should be well elaborated. Template analysis also requires attention to the environment and specific experience of the interview candidate to ensure that themes do not apply across a group of diverse participants. Furthermore, the method provides little guidance on how to move from coded data to its final interpretation, which means that the researcher should engage in more interpretative thinking before moving on to the outcomes.[24] In addition to the limitations of the method of analysis, a general limitation of this study is that it only examines the state of play in the Dutch financial sector, meaning that results may not be applicable to financial sectors/institutions in other countries. Still, the cross-border operations of Dutch financial institutions and their relative technological maturity make it likely that the findings are relevant to the broader European financial sector.

1.5. Structure

Not counting the introduction and conclusion, this thesis is divided into four substantive chapters. To gain more insight into the specific cyber risks faced by AI systems, the first chapter reviews recent literature on the phenomenon, establishing a generic AI lifecycle and asset taxonomy, a comprehensive overview of possible asset vulnerabilities, and potential threats that target said vulnerabilities. The second chapter examines current academic literature on cyber risk management in general and cyber risk management for AI systems in particular. The third chapter reports the findings and observations gained through the semi-structured interviews. The last and final chapter then provides a cross-

[23] *See* King *et al*, *supra* note 19, at 85-94.
[24] Ibid.

comparison between the academic literature and the interview observations and findings, analysing whether there are gaps and potential areas for improvement in future research and/or cyber risk management practices in the Dutch financial sector. The study concludes with an overview of the main findings and a brief discussion of future research directions.

2. Cyber risks for AI systems

To gain insight into cyber risk management methods and practices for AI systems, it is necessary to first understand the cyber risks that AI systems face. As such, following the structure of the ISO27005, which states that risks emerge when "threats abuse vulnerabilities of assets to generate harm for the organisation, this first substantive chapter examines general AI system assets, possible asset vulnerabilities, and potential threats to these assets.[25]

2.1. AI asset taxonomy

While AI assets are ultimately specific to the system in use, making it important for an organisation utilising an AI system to identify the assets used, for the purposes of this chapter, a generic AI asset taxonomy is established that can reasonably be assumed to address most current AI systems. To understand the asset taxonomy, it is important to first gain insight into what an average AI lifecycle looks like. For this, the generic AI lifecycle model as composed by the European Union Agency for Cybersecurity (ENISA) in their December 2020 AI threat landscape report is briefly examined.[26]

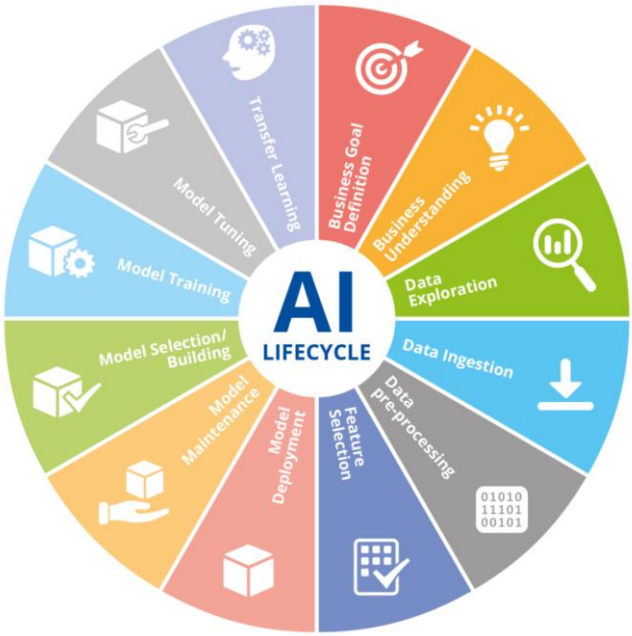

Figure 1: Generic AI lifecycle.[27]

The AI lifecycle presented is composed of several interdependent stages that feed back into each other, resulting in actions related to model maintenance being utilised for input in the data pre-processing stage, and model maintenance being utilised for feature selection.[28] These stages, like most AI lifecycle models, can be divided into a design, development, and deployment phase, with each phase requiring specific expertise and assets.[29] The lifecycle stages can be defined as follows:[30]

[25] (https://www.iso.org/standard/75281.html), last visited (29-10-2022).
[26] ENISA, AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence, December 2020, p. 13.
[27] Id, p. 58.
[28] Id, p. 14.
[29] D. De Silva and D. Alahakoon, An artificial intelligence life cycle: From conception to production, Patterns 3 (2022), pp. 1-10; M. Haakman, L. Cruz, H. Huijgens and A. van Deursen, AI lifecycle models need to be revised: An exploratory study in Fintech, Empirical Software Engineering (2021), pp. 4-23.
[30] *See* ENISA, *supra* note 26, at 16-21; *See* De Silva *et al*, *supra* note 29, at 1-10.

2.1.1. Design phase

Generally, in the design phase, the organisation defines a problem and conceptualises a solution based on the resources at hand. Drawing on the ENISA lifecycle, the design phase includes the following five stages: 1) Business goal definition, composed of investigating the business context, identifying and formulating the problem and the business purpose of the AI solution, including choosing the relevant AI model type; 2) data ingestion, here the required input data is identified and prepared, focusing on the data structure rather than the actual data; 3) data exploration, where the structure is populated with actual data; 4) data pre-processing, in which, through data conversion and other actions, it is ensured that the acquired data can be loaded into the model without comprising accuracy, informational value, and data quality, and 5) feature selection, where the data set dimensions are determined, ensuring that only those data components are selected that are meaningful for the AI model, reducing computational cost and increasing model accuracy.[31]

2.1.2. Development phase

The development phase broadly consists of three stages: model selection / building; model training, and model tuning. In the model selection / building stage, a suitable AI model is built. AI models can generally be divided into supervised, unsupervised and reinforcement learning models. It is common practice to begin the development process with a simple algorithm, and default parameters and architecture to ensure that the approach can be validated.[32] In the model training stage, the training algorithm is applied with the right parameters to modify the chosen model and validate model training. This mostly applies to a machine learning model in combination with supervised learning techniques. The model tuning stage involves the application of model adaption to the parameters of the training model using a data set for validation purposes.[33]

2.1.3. Deployment phase

The deployment phase includes all stages involved in implementing the AI system in the organisation. Drawing on the generic AI lifecycle it involves transfer learning, where a pre-tuned and trained model is externally sourced and further trained to improve accuracy.[34] In the case of sourcing a model, training, testing, and tuning are still required. Furthermore, this phase includes model deployment, this stage is also known as model serving, scoring, and production, it involves deploying the model and connecting production data flows, making it available to users. There is also model maintenance, where interference results and input data are monitored to detect an impact on model accuracy so that the model can be retrained when necessary. Last, this phase includes business understanding, which describes the stage in which the organisation that deployed the model type gains insight into its impact and undertakes organisational actions to maximise success.[35]

It is important to understand that next to these lifecycle stages, which are based on the ENISA model, it is possible to discern many other stages depending on the scope and perspective of the author(s). For instance, De Silva and Alahakoon's AI lifecycle model is composed of 19 distinct stages that include a stage dedicated to the review of data and AI ethics and a stage dedicated to the process of external data acquisition.[36] Furthermore, there are many other lifecycle models, such as CRISP-DM and TDSP,

[31] *See* De Silva *et al*, *supra* note 29, at 1-10.
[32] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Pearson Higher Education (2021), pp. 690-700.
[33] *See* De Silva *et al*, *supra* note 29, at 1-10.
[34] K. Weiss, T.M. Khoshgoftaar, and D. Wang, A survey of transfer learning. J Big Data 3, 9 (2016), pp. 1-4.
[35] *See* De Silva *et al*, *supra* note 29, at 1-10.
[36] Id, p. 4.

that are widely adopted. However, these models do not provide the granularity to further examine assets per lifecycle stage.[37]

Building on the structure of the lifecycle model above, it is possible to identify various assets per lifecycle stage, with assets being defined as "…those that that are crucial to meet the needs for which they are being used."[38] Following this definition, identified assets include the processes, actors and stakeholders engaged in the design, development, and deployment of AI-powered systems. This provides a long list of assets, which can be broadly divided into six categories, namely: 1) Data; 2) model; 3) actors; 4) processes; 5) environment/tools, and 6) artefacts, resulting in the following overview:

Figure 2: AI asset taxonomy[39]

The next paragraphs briefly go over the contents of these six asset categories. First, one of the most important components or assets of AI-powered systems is data. This can be raw data, which is any type of non-transformed data or non-enriched information that can be utilised for analysis purposes, a labelled data set, which is a data set tagged with informative labels, and training data, which is data

[37] *See* Haakman *et al*, *supra* note 29, at 4-23; (https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview), last visited (29-10-2022); (https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview), last visited (29-10-2022).
[38] *See* ENISA, *supra* note 26, at 22.
[39] Ibid.

used for training the chosen model. These assets all have a direct relation to data and are used in almost every stage of the AI lifecycle. Second, models can be considered the second major asset category of AI-powered systems. This includes data pre-processing algorithms, used to clean, integrate, and transform data to improve quality, and training algorithms, which are procedures for adjusting AI model parameters. Just like data, model assets are used throughout the whole AI lifecycle. The third category, actor assets, is more generic in the sense that they constitute an important asset category in all information technologies. This category includes actors and stakeholders, such as data owners or business owners, data scientists and engineers, as well as the end users. All actors play an essential role in the functioning of the AI-powered system and can be found in the whole AI lifecycle.[40]

The fourth asset category includes all actions, methods and techniques needed to ensure the proper design, development, and implementation of AI-powered systems. This includes processes like data ingestion, which relates to the transportation of data from multiple sources to create data points, and data exploration, which can be considered an asset as well as a stage in the AI lifecycle proper. Environment/tools, which is the fifth asset category, include all hardware and software on which an AI-powered system depends for its functioning. Assets include communication networks and protocols, processors providing computational power, and any cloud infrastructure in place. Last, the sixth asset category relates to all tangible by-products of the AI lifecycle phases, most relating to software documentation. Assets include access control lists, model architecture and data governance policies.[41]

2.2. AI vulnerabilities

All information technology (IT) assets or components have specific vulnerabilities that ultimately translate into cyber risk. The previous sub-chapter examined assets and asset categories that can be found in each stage or phase in a generic AI lifecycle. Using recent academic research, relevant vulnerabilities for each of the six asset categories (data; model; actors; processes; environment/tools, and artefacts) are identified.[42] Furthermore, it is worth noting that vulnerabilities related to the technical supporting infrastructure enabling the functioning of the AI-powered system are not examined since they are too generic. Moreover, some vulnerabilities are relevant for multiple asset categories.

2.2.1. Vulnerabilities of data assets

Data assets are vulnerable to manipulation and any kind of unintended modification or disclosure, specifically data assets can be vulnerable due to, inter alia: Inadequate data management, resulting in the confidentiality, integrity and availability (CIA) of the data being at risk; annotation issues to label data related to spurious labelled data or labels being deleted or omitted; using uncontrolled data, resulting in data being inconsistent; too little data decreasing robustness to poisoning; no data poisoning detection in the training dataset; unprotected sensitive training or test data; existing biases in the model or data; using unsafe data or models relevant in case, e.g., data or models are externally sourced;[43] incorrect tensor property values: relevant for software security vulnerabilities in ML libraries, this relates to threads or programmes not maintaining tensors (a mathematical object describing a multilinear relationship between sets of mathematical objects) adequately; incorrect type conversion: vulnerabilities related to incorrect conversion of data types; using improper data types: vulnerabilities related to incorrect handling of data types, and numerical precision errors: vulnerabilities related to defined variables or tensors within an improperly defined range.[44]

---

[40] Id, pp. 32-35.
[41] Id, pp. 35-42.
[42] ENISA, Securing Machine Learning Algorithms, December 2021, pp. 3-16; N.S. Harzevili, J. Shin, J. Wang, and S. Wang, 'Characterizing and Understanding Software Security Vulnerabilities in Machine Learning Libraries', Cornell University, 2022; N. Bouacida and P. Mohaptra, Vulnerabilities in Federated Learning, IEEE Access Volume 9, 2021; Y. He, G. Meng, K. Chen, X. Hu, and J. He, 'Towards Security Threats of Deep Learning Systems: A Survey', IEEE Transactions on Software Engineering 2020.
[43] *See* ENISA, *supra* note 42, at 15-16.
[44] *See* Harzevili *et al*, *supra* note 42, at 3-6.

2.2.2. Vulnerabilities of model assets

Model asset vulnerabilities mostly emanate from cyber risk and more specifically cyber security not receiving due consideration in the design, development, and deployment phase of the AI lifecycle. These vulnerabilities mostly relate to using inherently insecure models and include: Inadequate consideration of evasion attacks in model design and deployment; utilising a known model, allowing adversaries to study it; model output providing too much information on the workings of the model; model output allowing sensitive information to be retrieved; existing biases in the model or data; specific model vulnerabilities to poisoning attack; inadequate implementation of access protection mechanisms for model components;[45] stack or buffer size issues: vulnerabilities related to inappropriate size definition for stacks or buffers; out of bound read: vulnerabilities related to reading information from incorrect memory locations; improper memory management: vulnerabilities related to the development phase, in which memory is incorrectly handled, e.g., inadequate use of memory release statement, and invalid memory access: vulnerabilities related to model processes accessing memory locations filled with null values.[46]

2.2.3. Vulnerabilities of actor assets

These vulnerabilities relate to organisational or human factors in the lifecycle stages. Actor and stakeholder asset vulnerabilities include: Third-party service provider security concerns, e.g., software supply chain vulnerabilities; lacking organisational cyber security awareness; inadequate documentation on the AI-powered system; inadequate consideration of cyber risk management and cyber security in the design, development, and deployment phases of the lifecycle, and lacking explainability and traceability of decisions taken.[47]

2.2.4. Vulnerabilities of process assets

Process asset vulnerabilities relate to insecure or deficient methods and actions in the phases of the AI lifecycle. These vulnerabilities include: The presence of unidentified disclosure scenarios; using vulnerable components in the lifecycle; unprotected sensitive data in the training or test environment; inadequate access rights management, and lacking security processes and procedures.[48]

2.2.5. Vulnerabilities of environment/tool assets

Any environment/tool the AI-powered system depends on has vulnerabilities, such as: Inadequate access rights management; inadequate data management; deficient or absent access protection mechanisms for model components, and third-party service provider security concerns.[49]

2.2.6. Vulnerabilities of artefact assets

Vulnerabilities related to artefacts include: The AI-powered system not being compliant with regulations; inadequate access rights management; deficient or absent access protection mechanisms for model components, and vulnerabilities resulting from poor integration of AI-powered system specificities and existing risk or security policies.[50]

---

[45] *See* ENISA, *supra* note 42, at 15-16.
[46] *See* Harzevili *et al*, *supra* note 42, at 3-6.
[47] *See* ENISA, *supra* note 42, at 15-16.
[48] M. Fredrikson, S. Jha, and T. Ristenpart, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security 2015, pp. 1322-1324; R. Shokri, M. Stronati, C. Song, and V. Shmatikov, Membership Inference Attacks Against Machine Learning Models, In the proceedings of the IEEE Symposium on Security and Privacy 2017, pp. 1-10.
[49] Y. Ji, Z. Zhang, S. Ji, X. Luo, and T. Wang, Model-Reuse Attacks on Deep Learning Systems, Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security 2018, pp. 1-4; Bundesamt für Sicherheit in der Informationstechnik, Towards Auditable AI Systems: Current status and future directions (2021), pp. 12-13.
[50] Ibid.

Next to these vulnerabilities, it is also possible to discern several vulnerabilities related to AI-specific ecosystems such as federated learning (FL), which is a decentralised training paradigm enabling distributed clients to train an AI model without sharing data, promising, e.g., enhanced privacy.[51] Since FL is increasingly being experimented with, and related vulnerabilities being AI-specific, it is worth it to briefly examine this phenomenon.

2.2.7. Vulnerabilities of AI-specific ecosystems (federated learning)

Vulnerabilities related to FL AI systems include those related to communication in FL, namely homomorphic encryption (HE), which is often used in FL to protect client data and HE has vulnerabilities of itself that can be exploited. In addition, FL must deal with bottlenecks related to internet connections operating at lower speeds than internal server communication links. This can result in clients dropping out, resulting in unwanted bias. Furthermore, there are specific vulnerabilities related to compromised clients in FL: since FL is dependent on distributed clients training the AI model, any client can observe the global model states and can contribute through updates. This provides opportunities for tampering with the training process. In addition, there are vulnerabilities related to the use of the aggregation algorithm in FL: this algorithm coordinates the global parameter learning through detection and discarding of abnormalities. Due to its central function, incorrect configuration of the algorithm poses a major vulnerability. Last, there are vulnerabilities related to the distributed nature of FL: distributed training has inherent vulnerabilities to colluding or distributed attacks against updates to the global model.[52]

2.3. AI threats and their potential impact

Following ISO27005, threats to AI-powered systems are defined as those threats that target AI vulnerabilities to generate harm for the financial institution, either maliciously or non-maliciously through negligence. Furthermore, threat actors are not examined but rather the types of attacks that abuse vulnerabilities identified in the previous section are analysed as well as the harm they generate in terms of their potential negative impact on the CIA of the AI-powered system.

2.3.1. Threats to data assets

There are several threats to data assets, with some of the most prominent being data poisoning attacks: these are types of attacks that seek to decrease the predictive accuracy of the model or create a backdoor through the pollution of training data. Moreover, contamination of the model occurs before the training phase and is very difficult to extract. Data poisoning attacks primarily impact integrity and availability.[53] Another important threat is the introduction of selection bias: raw data can be tampered with to introduce selection bias to steer output in an attacker's chosen direction, also known as an adversarial perturbation, adversely affecting data inference, and impacting integrity and availability.[54] Last, another prominent threat is mishandling of statistical data: if the model does not allocate proper rewards or shares, any results might suffer from a skewed distribution, impacting confidentiality and availability.[55]

2.3.2. Threats to model assets

Generally, there are three major threats to model assets. These are: 1) model poisoning attacks, which refers to models being tampered with or replaced, impacting integrity and availability. For example, a model might be replaced in the context of AI-as-a-Service through the exploitation of weaknesses in the

[51] *See* Bouacida *et al*, *supra* note 42, at 63229.
[52] Id, pp. 63232-63233.
[53] *See* He *et al*, *supra* note 42, at 10-11; C. Anley, Practical Attacks on Machine Learning Systems, NCC Group (2022), pp. 26-27.
[54] S. Datta, N. Shadbolt, Backdoors Stuck At The Frontdoor: Multi-Agent Backdoor Attacks That Backfire, arXiv:2201.122211v1, pp. 1-8.
[55] *See* ENISA, *supra* note 26, at 46-51.

cloud services;[56] 2) model inversion attacks, also known as training data extraction: these attacks leverage model information flows in the training process to infer data, impacting confidentiality, and 3) model extraction attacks, also known as model stealing attacks: these kinds of attacks duplicate an AI model through an application programming interface, impacting confidentiality of the original.[57]

2.3.3. Threats to actor assets

Three noteworthy threats to actor assets include: 1) misconfiguration or mishandling of AI-powered systems, where actors/stakeholders can expose data or model functioning through negligence, impacting confidentiality; 2) lacking data protection of third parties, involving third-party service providers are often involved in the provision or processing of data, which can result in data being exposed, impacting confidentiality, and 3) compromise of data brokers/providers: compromised data brokers or providers can influence the model learning process, impacting integrity and availability.[58]

2.3.4. Threats to process assets

There are various threats to process assets, such as manipulation of model tuning, which involves the manipulation of parameters by adversaries to change AI-powered system behaviour, impacting integrity and availability, scarce data, relating to the situation where by not having reliable access to data a model's viability or results can be compromised, impacting availability, and data tampering, which is the deliberate or unintentional manipulation and exposing of data by taking advantage of a poorly designed or implemented process, impacting integrity and availability.[59]

2.3.5. Threats to environment/tools assets

There are various other threats to environment/tools assets such as DDoS threats. However, since they relate to technical supporting infrastructures they are not in the scope of this research. One non-technical threat worth mentioning is service level agreement breach: in case the AI-powered system is dependent on third parties, a breach of agreed-upon service levels can result in performance degradation, impacting availability.[60]

2.3.6. Threats to artefact assets

Threats to artefact assets are numerous, including corruption of data indexes, where data index contents can become corrupted through deliberate attack or unintentional system or network failure, impacting integrity and availability, and poor resource planning, in which inadequate computational resources can result in the AI-powered system not functioning well, impacting integrity and availability.[61]

2.3.6. Threats to AI-specific ecosystems assets

Federated learning systems have to deal with three macro threats, namely: Man-in-the-Middle attacks: in the case of FL, data and models exchanged between clients can be intercepted to replace them with malicious content, impacting confidentiality and integrity; dropout of clients: operational failures such as network issues in FL can result in clients dropping out, depriving the model of data and impacting availability, and non-robust aggregation: aggregation algorithm in FL with weak defence mechanisms or inappropriate reweighing schemes can result in abnormal behaviour of the global model, impacting integrity.[62]

---

[56] *See* National Security Commission on Artificial Intelligence (NSCAI), *supra* note 3, at 45-52.
[57] *See* He *et al*, *supra* note 42, at 4-7; *See* Anley, *supra* note 53, at 25-26, 29.
[58] *See* ENISA, *supra* note 26, at 47-52; T. Gu, B. Dolan-Gavitt, and S. Garg, BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, arXiv:1708.06733 (2019), pp. 10-12.
[59] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain, arXiv:2007.02407 (2021), pp. 1-14.
[60] *See* ENISA, *supra* note 26, at 52.
[61] Id, pp. 53-54.
[62] *See* Bouacida *et al*, *supra* note 42, at 63234.

2.4. Summary

This chapter provided an overview of cyber risks to AI systems, examining cyber risk through its constituent components of assets, vulnerabilities, and threats. In general, whenever there is a new technology, there are also new threats such as attack techniques that exploit vulnerabilities, and AI does not escape this rule. While all threat categories examined in this chapter should be considered when managing cyber risks for AI-powered systems, data-level and model-level risks are arguably the most material, in which model stealing, poisoning and adversarial attack threats play an important role.[63] The insights into cyber risks obtained in this chapter provide the foundations for further exploration of relevant cyber risk management practice in general and cyber risk management for AI systems in particular. Furthermore, while not all examined risks are only applicable to AI systems, their specific assets, their vulnerabilities, and relevant threats mean that specific trade-offs need to be made that may not be applicable to other more established information technologies. Moreover, the potential interconnectivity and complexity of AI-powered systems as well as their reliance on greater data volumes can result in a larger organisational attack surface. Taking these specificities into account, the following chapter investigates relevant cyber risk management practices.

[63] EBA, EBA Report on Big Data and Advanced Analytics (2020), p. 41.

3. Theory on cyber risk management for AI systems

The previous chapter established an overview of cyber risks to AI systems, constructed from its risk aspects in line with ISO27005, i.e., AI components, components' vulnerabilities, and the various threats to these vulnerabilities. This chapter explores recent cyber risk management practices put forward by the research community applicable to the identified cyber risks for AI-powered systems.

3.1. Cyber risk management in general

Cyber risk can be defined in numerous ways, but for the purposes of this study, the following comprehensive definition is used:

"Cyber risk is an operational risk associated with performance of activities in the cyberspace, threatening information assets, ICT resources and technological assets, which may cause material damage to tangible and intangible assets of an organisation, business interruption or reputational harm. The term 'cyber risk' also includes physical threats to the ICT resources within organisation."[64]

This definition includes the most important features of cyber risk, namely that: 1) cyber risk is placed among operational risks (as accepted by the academic community); 2) the relevancy of cyberspace as source of cyber risk; 3) objects exposed to cyber risk include information and technological resources, as well as IT resources; 4) cyber risk can manifest in a single computer resource or in a computer network, and 5) the potential impact of cyber risk can result in property damage (both tangible and intangible), disruption of operations or damage to reputation.[65] While this study acknowledges that cyberspace is one of the most relevant sources of cyber risk, non-cyberspace sources of risk, such as physical threats, are not placed out of scope.

With the term cyber risk management, this study means that a risk management process is used to manage cyber risk. Due to risk management processes being highly divergent between sectors and fields, and to prevent semantic confusion, in line with the generic ISO31000, this study follows the basic risk management process of Michael Elliott, with a focus on cyber risks:[66]

1. Environmental scanning
2. Risk identification
3. Risk analysis
4. Risk treatment (avoidance; mitigation; transfer; retention, and exploitation)
5. Risk monitoring and process review

Due to this thesis's limited scope, the first and last part of the risk management process are not covered. Environmental scanning relates to the incorporation of risk management in the overarching corporate governance and its alignment with the objectives and risk appetite of the organisation, and risk monitoring and process review relates to third line (internal audit) and senior management reporting and steering processes.[67] The following paragraphs provide an overview of scholarly contributions to risk management practices regarding the risk identification, risk analysis and risk treatment processes.

3.1.1. Cyber risk identification

Towards the second half of the 1990s the cyber security research community became aware of the need for universal frameworks to ensure consistency and integrity in the identification and categorisation of cyber-attacks. Howard and Longstaff were one of the first in consolidating relevant discussions in a useable framework for risk identification that used common language largely in line with what are now

[64] *See* Strupczewski, *supra* note 13, at 6.
[65] Ibid.
[66] M. Eling, M. McShane, T. Nguyen, Cyber risk management: History and future research directions, Risk Management and Insurance Review (2021), p. 96; M. W. Elliott, Risk in an evolving world, The Institutes (2019), pp. 1.23-1.27; (https://www.iso.org/iso-31000-risk-management.html), last visited (20-12-2022).
[67] *See* Eling *et al*, *supra* note 66, at 96-97; IIA, The IIA's Three Lines Model: An update of the Three Lines of Defense (2020), p. 3.

considered standard risk identification steps (assets, vulnerabilities, and threats) in the risk management process. In their framework, an incident involves an attacker, such as hackers or professional criminals, reaching their objectives, such as financial or political gain, through an attack or attacks. An attack then consists of a tool like a script or programme that targets a vulnerability in either the design, implementation, or configuration of a system, resulting in an event that leads to an unauthorised outcome. Last, an event contains both an action, such as a probe or scan, and a target like a process. In more modern risk identification frameworks, targets would be considered assets.[68]

Another way to identify and classify cyber-attacks is to examine their impact on the CIA of information assets. For instance, a denial-of-service attack mainly affects availability whereas a phishing attack primarily impacts confidentiality.[69] The cyber risk identification framework of Howard and Longstaff, and classification through impact on the CIA of information assets are widespread methods to identify and categorise cyber-attacks. More modern methods for risk identification are more technology-driven, using big data and machine learning analysis techniques to detect anomalies posing a threat.[70] Furthermore, recently there has been an academic focus on more proactive approaches to cyber risk identification, including the use of honeypots to attract and analyse attacks before they become mainstream.[71] Still, there are gaps in cyber risk identification research regarding recent developments, such as cloud computing, the Internet-of-Things and AI. These technologies have resulted in increased interconnectivity between new and existing assets, giving rise to fresh vulnerabilities and threats, further increasing the potential attack surface of organisations.

3.1.2. Cyber risk analysis

The process of cyber risk analysis relates to the investigation of the likelihood and financial or operational impact of a cyber-attack on an organisation. Cyber risk analysis research can roughly be split up between research focused on the likelihood or probability of an incident, and research focused on the specific impact of an incident. Concerning research on probability, scholars find that organisations with certain characteristics are more likely to experience a cyber incident. However, the evidence regarding what characteristics increase or decrease the likelihood of an incident greatly differs due to researched time periods and sampling criteria.[72] For instance, research on organisational size being a factor in cyber incident likelihood shows that smaller organisations' incident chance is greater than larger organisations, with a possible explanation being that larger organisations have more information security resources.[73] Other factors include organisations' age, value, capital expenditures, research & development spending, and growth opportunities. Recent research on probability focuses on the application of new statistical models, adversarial risk analysis and the application of machine learning techniques, such as applying Bayesian belief networks and developing risk scoring tools assessing attack probabilities based on continuously updated software vulnerabilities and assets at risk.[74]

Concerning cyber risk analysis research focusing on impact, there have been numerous studies examining the consequences of a cyber incident on shareholder value, with studies often contradicting each other. Less research has been carried out on the effect of cyber incidents on the long-term performance of organisations, but recent studies do show agreement that cyber incidents generally result in additional costs for the affected organisation, impacting financial performance and productivity.

---

[68] J. D. Howard & T. A. Longstaff, A common language for computer security incidents, Sandia National Labs (1998) (No. SAND98-8667), p. 16; *See* Eling *et al*, *supra* note 66, at 99-101.
[69] *See* Eling *et al*, *supra* note 66, at 100.
[70] Q. Lin, S. Verwer, S. Adepu and A. Mathur, TABOR: A Graphical Model-based Approach for Anomaly Detection in Industrial Control Systems, ACM Asia Conference on Computer and Communications Security (2018), pp. 525-534.
[71] A. Marotta & M. McShane, Integrating a proactive technique into a holistic cyber risk management approach. Risk Management and Insurance Review (2018), pp. 435–452; *See* Eling *et al*, *supra* note 66, at 100.
[72] *See* Eling *et al*, *supra* note 66, at 101-103.
[73] C. Lending, K. Minnick and P. J. Schorno, Corporate governance, social responsibility, and data breaches, Financial Review (2018), pp. 413–455; *See* Eling *et al*, *supra* note 66, at 101-103.
[74] *See* Eling *et al*, *supra* note 66, at 102.

Furthermore, there is evidence that affected organisations increase risk management investment and shift towards more risk-averse policies. In addition, research on the impact of cyber incidents shows that cyber incidents can result in decreased consumer trust and spending. Last, there has also been research on possible spill over effects of cyber incidents, impacting third-party providers and peers of the affected organisation.[75] Still, spill over effects is difficult to measure and it remains unclear to what extent the possibility of spill over impacts the cyber risk management decisions of organisations.

3.1.3. Cyber risk treatment

There are several ways to treat cyber risk, namely through avoidance (if possible) and mitigation when benefits are greater than costs, and through transferring or retention of risk if in line with the organisation's risk appetite. While dependence on IT makes risk avoidance in most cases an unrealistic type of risk treatment, most organisations have numerous risk mitigation controls in place. Research often uses the Parkerian hexad to identify risk mitigation opportunities. This hexad consists of the CIA triad plus authenticity, possession or control, and utility. Academic research on technical mitigation often categorises treatment opportunities based on relevant security issues, such as secure communication and security management, and applicable techniques, such as cryptographic techniques and methods of authentication.[76] In addition, there is numerous academic research on the economics of cyber risk treatment, such as research carried out by Gordon and Loeb, who argued that security investment should not exceed 37% of the predicted loss.[77]

Concerning cyber risk treatment research on risk transfer, the academic discussion has mainly revolved around the use of insurance as risk transfer tool, with the main insurance challenges being a lack of data for pricing, adverse selection, moral hazard, information asymmetries, and highly interrelated losses.[78] Another popular risk transfer option in practice is outsourcing the assets or processes at risk. However, academic research on outsourcing as risk transfer tool is limited. Regarding risk retention, research has been carried out showing that breached organisations have more cash holdings than non-affected organisations.[79] However, there seems to be a lack of research attention to optimal risk retention practices.

3.2. Cyber risk management for AI systems

The risk management process, consisting of risk identification, analysis and treatment, is also relevant to the cyber risk management for AI systems. However, as already mentioned, it is important to consider that the specific components of AI systems, their vulnerabilities and relevant threats mean that specific trade-offs need to be made that may not be applicable to other more established information technologies. Having identified recent general cyber risk management practices developed by the research community in the previous sub-chapter, starting with cyber risk identification, the following paragraphs investigate recent research on risk management practices specifically applicable to cyber risk management for AI systems.

3.2.1. Cyber risk identification for AI systems

As examined in the previous sub-chapter, risk identification is broadly performed in two manners: 1) by subdividing risk into its constituent components, or 2) by analysing risk impact on the CIA of a specific information asset. Exploring recent academic articles on risks for AI systems shows that risk identification for AI systems is mostly carried out by the same two means, with specific attention to the AI lifecycle and its data and model components. Regarding risk identification and classification through

[75] Id, p. 104.
[76] Id, p. 105.
[77] L.A. Gordon and M.P. Loeb, The Economics of Information Security Investment, ACM Transactions on Information and System Security (2002), pp. 438-457.
[78] *See* Eling *et al*, *supra* note 66, at 106-107.
[79] P. Garg, Cybersecurity breaches and cash holdings: Spillover effect, Financial Management (2020), pp. 503–519; *See* Eling *et al*, *supra* note 66, at 106-107.

subdivision into constituent components, Zhang et al conduct a comprehensive examination of risks for AI systems, identifying inherent risks in these systems that arise from different sources, defining, describing, explaining, and categorising each type of risk. They posit that risks in generic AI systems can be grouped into two broad categories, both inherent to the system itself: data-level risk, and model-level risk.[80]

Data-level risk originates from AI systems' dependence on data assets with specific vulnerabilities related to data bias, dataset shift, which is a shift in datasets related to its distribution, out-of-domain data, related to input falling outside of the defined problem domain, and adversarial attacks. Model-level risk pertains to model bias, misspecification, and uncertainty. Zhang et al argue that AI systems introduce various risks with unique characteristics that cannot be identified and managed through a traditional risk management framework. By focusing on the inherent risks stemming from AI systems' data and model assets for risk identification purposes Zhang et al hope to better address risks for AI systems to increase the trustworthiness in adopting and safeguarding their usage in high-risk environments.[81] While the data and model risk focus of Zhang et al allows for a strong defined approach to risk identification, they do not consider other sources of risk stemming from AI systems' other assets, vulnerabilities and threats.

A similar risk identification process for AI systems can be found in a recent article by Mauri and Damiani who researched threat modelling to AI systems using STRIDE, a well-known model first developed by Microsoft. Their starting point is a generic AI life cycle, which provides insight into the components or assets that are vulnerable to various threats, translating into numerous risks for AI systems. Their findings show that the difficulty inherent in this approach is that it is very ambitious to gain a complete or even adequate overview of the attack surface for AI systems. This is due to the AI-specific attack surface expanding along a new axis resultant from the multifaceted and dynamic nature of AI components and processes. Consequentially, the attack surface is very complex, and further mapping requires going through all steps of the AI system's specific life cycle and explaining all relevant security threats, a challenging process due to the numerous attack vectors. As such, Mauri and Damiani focus on six at-risk macro-categories of AI assets, namely: data, models, actors, processes, tools, and artefacts.[82]

Marui and Damiani then use Failure Mode and Effects Analysis (FMEA) to identify, prioritise and limit failure modes of these AI assets, which covers the risk identification, analysis, and mitigation process of risk management. They demonstrate that FMEA can be applied to AI systems by following a four-step guide: 1) Creating a function list per AI asset, whose content should be different for each asset in function; 2) specification of prerequisites for functions that can refer to functions of other identified assets, which provides the basis for creation of function networks; 3) identification of asset defects that potentially impair a function, i.e., the failure mode (FM). Each defect should be accompanied by causes and effects, and 4) using a threat-modelling methodology like STRIDE to map FMs to threats.[83]

In practice, this means that asset category owners, action managers and security analysts are interviewed to identify functions of AI systems, the assets' failure mode, and finally, the effects of the failure mode. For instance, for function identification purposes, an asset category owner could be asked what the primary purpose of the asset is or what it is supposed to do. For identification of FMs, the question could be posed on in what way the asset could fail in performing its intended function. Finally, to identify effects for FMs, an owner, manager or analyst could be asked what the consequence of failure is or whether failure could harm users or be a breach of applicable regulation.[84] Next to applying FMEA for risk identification purposes, Mauri and Damiani also examine the impact of cyber risks on the CIA

[80] *See* Zhang *et al*, *supra* note 8, at 2-8.
[81] Id, pp. 9-11.
[82] *See* Mauri *et al*, *supra* note 8, at 1-6.
[83] Id, p. 5.
[84] Id, pp. 5-6.

plus authenticity, non-repudiation and authorisation of AI system assets, providing a more intricate overview, which can be combined with a DREAD scorecard to prioritise risk.[85]

This supportive method involves examining a specific AI system asset, e.g., a training data stream asset, the potential damage that could result from a cyber risk, such as a spoofing attack resulting in backdoor generation and data substitution, potentially altering metrics and reporting, and the constraints in place to prevent damage from occurring. Resultant scenarios can then be scored on DREAD categories: damage, reproducibility, exploitability, affected users, and discoverability, allowing for risk prioritisation next to risk identification. In general, Mauri and Damiani find that risks for AI systems chiefly negatively impact performance and/or result in system misbehaviour, and/or privacy breach.[86] The potential downside of Mauri and Damiani's approach to cyber risk identification and prioritisation for AI systems is that while being comprehensive and well-structured it requires continuous updating of failure modes, system information itself and the latest research trends on weaknesses and vulnerabilities.[87] While this is also true for a lot of established cyber risk management practices, it is resource intensive.

3.2.2. Cyber risk analysis for AI systems

As discussed in the previous sub-chapter, cyber risk analysis research can be roughly split up into research focusing on likelihood and probability, and research focusing on impact. Jia and Zhang in their article on a risk analysis framework for AI systems, focus on the classic risk management literature rather than on the technical life-cycle model of AI. To measure the probability of risk, they target factors affecting the uncertainty of whether and how risk would happen, arguing that two factors are of special importance due to the characteristics of AI, namely the environment and the stakeholders involved. Jia and Zhang posit that the environmental factor is important because the technical capacity of AI is still very much in development and its application is not mature. Therefore, while risk analysis should probably focus on the AI system itself in case of a more developed and mature implementation of such a system, for now, the uncertainty and analysis of risk are mostly dependent on the environment, i.e., external factors. The second factor is important because the risk is specific regarding the stakeholders involved, the traceable causes and assumed responsibilities. In this sense, they posit that for determining the probability of risk to AI systems and risks emanating from AI, one should approach it from an ecosystem perspective.[88] For this purpose, general cyber risk practices are sufficient for determining probability.

To measure the impact of risks on AI systems and to prioritise mitigation efforts, it is also possible to use bug bars, which provide a taxonomy for classifying vulnerabilities or threats based on their characteristics and severity. In November 2019, the Microsoft AETHER Engineering Practices for AI Working Group published a bug bar for ranking AI threats, which is supposed to supplement the existing software development lifecycle bug bar.[89] Their bug bar provides practical guidance for threat modelling AI systems and their dependencies and focuses on intentional malicious behaviour specifically targeting AI systems. The severity of the indication in this bug bar is based on the operational impact the threat or risk has on the well-functioning of a generic AI system. For instance, adversarial perturbation, which is an attack where a query is modified in a manner to receive a desired response from a production-deployed system, is, except for a random classification type adversarial perturbation attack, classified as having a critical severity or impact on AI systems.[90] While such a bug bar does not consider the financial or operational impact of risk on the organisation as a whole, it does provide a good starting point for further cyber risk analysis for AI systems. The shortcomings of bug bars are that they do not

[85] Id, pp. 7-13.
[86] Id, pp. 7-13.
[87] Id, p. 17.
[88] *See* Jia *et al*, *supra* note 17, at 62-64.
[89] (https://learn.microsoft.com/en-us/security/engineering/bug-bar-aiml), last visited (15-10-2022).
[90] (https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml#1-adversarial-perturbation), last visited (15-10-2022).

take into account organisational circumstances or its environment, meaning that while being useful as a supportive method for cyber risk analysis, they should be complemented by other practices.

In their article on sources of risk in AI systems, Steimers and Schneider propose a risk management process for AI systems, consisting of five steps: 1) definition of risk acceptance criteria; 2) risk assessment, consisting of risk identification and risk analysis; 3) risk evaluation; 4) risk control, and 5) market observation. In this process, risk analysis as part of the risk assessment process can be further split up into the potential extent of the damage (impact) and the probability of occurrence. Following Steimers and Schneider their risk management process, impact, and probability for threats to and emanating from AI systems can be determined by following more general risk assessment processes, such as those drawn up in the ISO12100 and ISO14971, which state that risk is determined by following three factors: 1) hazard exposure; 2) occurrence of a hazardous event, and 3) possibility of avoiding or limiting the harm. They do note that risk analysis is difficult due to there being little experience in the development and the use of applications based on this new technology.[91] While it can be argued that any cyber risk analysis process for AI systems should be encapsulated in existing risk management practices for standardisation, adoption and ease of use purposes, general risk assessment processes do not consider the specific characteristics of AI systems, meaning that it is difficult to determine whether and how any relevant risks are adequately managed.

3.2.3. Cyber risk treatment for AI systems

Research into risk treatment for AI systems through risk avoidance, mitigation, risk transfer or retention is numerous. Of these risk treatment options; risk mitigation is most relevant in practice. Based on research of Steimers and Schneider, it is possible to discern three risk mitigation strategies for AI-powered systems, namely: 1) creation of an inherently safe design; 2) implementing safeguards, and 3) providing information for end users. They argue that the creation of an inherently safe AI system should always be attempted and when this is not possible, safeguards should be implemented, including informing end users of relevant cyber risks. For the creation of an inherently safe design, it is possible to follow the structure of ISO12100, a general ISO standard on the safety of machinery, which includes general principles for design, risk assessment and risk reduction. Based on this standard, technical measures for the creation of a safe design and relevant safeguards can broadly be based on four pillars: inherently safe design, safety reserves, safe failure, and safety-related protective measures. Steimers and Schneider argue that these pillars should also be considered when designing AI systems, taking into account their specific characteristics, such as their reliance on quality data.[92]

Concerning the first pillar, a good starting point for designing an inherently safe AI-powered system is to, where possible, use a simple AI model. This follows from AI systems performance relying to a significant extent on quality data and that complex models are not very transparent in the manner that it is difficult to understand the decision-making process, making it difficult to trace erroneous results or general malfunctions. Moreover, a model with a lower complexity allows for easier interpretability, permitting easier manual maintenance and error checking. Regarding the second pillar, safety reserves, just like in mechanical systems where a tolerance range should be determined for their operations, in AI systems it is important to consider their limits for reliable decision-making. As such, preferred models are those that can calculate a measure for the uncertainty of their prediction. This is also applicable to the third pillar, safe failure, if the measure of the uncertainty of the prediction is relatively high, the AI system may require human verification.[93]

For the fourth pillar, safety-related protective measures can be implemented in various manners, from the application of a quality-assurance process to external protection devices. Furthermore, in the design

[91] A. Steimers and M. Schneider, Sources of Risk of AI systems, International Journal of Environmental Research and Public Health (2022), pp. 3-5.
[92] Id, p. 6.
[93] Id, pp. 6-7.

of AI systems, general security practices should be considered, such as secure software development lifecycle practices or more specific standards like the ISO/IEC TR 24028:2020 that provide an overview of possible AI system engineering mitigation techniques and methods.[94] In addition, it is also possible to consider mitigation of AI-specific risks from a security perspective, for which recent research has mostly focused on adversarial attacks on AI systems, in which an attacker manipulates a system to cause it to malfunction, change expected output or infer information. Here, risk mitigation can be implemented on a hardware and software level. On a hardware level, AI systems can be made more resilient against modification of inputs by employing a local, non-cloud AI model, directly connected to sensors.[95]

On a software level, AI systems can be made more resilient against adversarial attacks through AI-specific hardening, robustification, testing and verification techniques. For example, adversarial examples could be introduced during training, providing the model with prior information about expected output.[96] In addition, when examining the bug bar of the Microsoft AETHER Engineering Practices for AI Working Group, most AI-specific risk mitigation strategies involve more generic mitigation controls, such as implementation of strong access control, rate-limiting queries, input validation techniques for model inversion attacks, minimisation or obfuscation of details returned in prediction APIs to mitigate model stealing attacks, and minimisation of 3rd party dependencies for models and data (where possible) for mitigation of AI supply chain attacks.[97]

Hardware and software level safeguards can be combined to mitigate the potential impact of specific risks for AI-powered systems, such as, inter alia, data risks emanating from data poisoning threats, introduction of selection bias/adversarial perturbation threats, and mishandling of statistical data, and model risks proceeding from model poisoning attacks, model inversion, and model extraction threats. For example, measures against data poisoning include the implementation of reasonable supply-chain checks on any training data used, and not allowing third-party service providers to modify training data. For adversarial perturbation, which consists of a broad class of attacks, mitigations include implementing a model training regime that results in models being more robust against these types of threats. Moreover, in the case of web-accessible models under control, mitigating measures include the implementation of authentication and rate-limiting, which may make an attack traceable and slower, and the configuration of alerts on unusual clusters of intensive use of the system.[98]

Continuing from the above, mitigations against mishandling of statistical data, which can result in unintended bias, include the implementation of an appropriate bias risk management process as part of the development phase.[99] For model risks, safeguarding/mitigating measures for model inversion include the implementation of a training regime that results in a model becoming more robust against these types of (privacy) attacks, by implementing rate limits and authentication, and ensuring that sensitive data is not present in the model. Mitigations against model extraction include ensuring that filesystems backing the model are secure and that access rights are adequately reviewed. Moreover, these threats can be mitigated through configuration of alerts on unusual clusters of intensive use of the system.[100]

---

[94] *See* Steimers *et al*, *supra* note 91, at 7; NPR-ISO/IEC TR 24028, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence (2020), pp. 23-33.
[95] *See* Steimers *et al*, *supra* note 91, at 21-22.
[96] Id, p. 22.
[97] (https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml?source=recommendations#3-model-inversion-attacks), last visited (16-10-2022).
[98] *See* Anley, *supra* note 53, at 26-27.
[99] M. Seng Ah Lee, J. Sing, Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle, Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021), pp. 1-6.
[100] *See* Anley, *supra* note 53, at 28-29.

It is also worth mentioning that there are some emerging best practices concerning cyber risk mitigation for AI that could guide organisations in their internal discussions regarding AI risk. The AI/ML Risk and Security (AIRS) working group, composed of several prominent AI, risk and cyber security professionals employed in the US financial sector, have summarised some of these best practices. Most relevant are the implementation of a robust oversight and monitoring process to validate system outputs, thresholds, and other system aspects that could maintain overall accuracy and efficiency. A starting point for the creation of such an oversight process could be for an organisation to take stock of all their AI systems, their specific uses, techniques used, names of all stakeholders such as developers, and a current overview of relevant risk ratings. Concerning set up of a monitoring process, accuracy drift can be mitigated by implementing a drift detection function that monitors data received by the model in production and estimates model accuracy. Moreover, data drift can be mitigated by establishing a monitoring function that assesses whether input data deviates from the model's training data. Other best practices include only sharing minimal information on model working to mitigate malicious actors taking advantage of available information, maintaining the privacy of training data with differential privacy, which adds random noise to a dataset, and the use of watermarking by training the AI system to produce unique outputs for certain inputs, allowing for identification of successful model extraction attacks.[101]

In addition to the more specific safeguarding/mitigating measures examined in the above paragraphs, financial institutions should also take into account supervisory and regulatory concerns regarding cyber risk mitigation for AI-powered systems. While regulatory risk is outside of the scope of this study, from a prudential perspective, the soundness of core systems is of primary concern, which overlaps with cyber risk management for AI. Regarding robustness of AI-powered systems, De Nederlandsche Bank (DNB) has formulated five aspects that should be considered, namely: 1) ensuring compliance regulatory obligations; 2) mitigating prudential risks in the development and use of AI; 3) paying attention to mitigation of model risk for crucial AI systems; 4) safeguarding and improving the quality of data used by AI, and 5) being in control of procured or outsourced AI applications.[102] These aspects can be operationalised into several practices, such as: 1) regulatory compliance is implemented in the design of the AI system and continuity of operations is ensured through fall-back plans; 2) among others, boundaries are set to constrain model outcomes and models are periodically retrained and recalibrated; 3) inter alia, a review process is implemented to deal with erroneous outcomes; 4) minimal requirements concerning data quality are defined and upheld, and 5) a robust AI supply chain risk management process should be implemented.[103]

3.3. Summary

This chapter provided an overview of academic research on cyber risk management in general and cyber risk management for AI-powered systems in particular. Following the generic risk management process structure risk identification, risk analysis, and risk mitigation, it is possible to discern several practices that are relevant and applicable to those organisations looking for manners to manage cyber risks for their AI systems. For cyber risk identification, organisations could opt to focus on the two, arguably, main inherent risk areas for AI-powered systems, data-level risk, and model-level risk. This allows for a strongly defined approach but could result in overlooking material risks emanating from AI systems' non-data or model-related assets, such as artefacts and stakeholders. Another approach to cyber risk identification for AI systems is the use of FMEA to identify and prioritise risks in which STRIDE and DREAD can be used to structure and categorise risk components and potential impact. The downside of this approach is that it is a resource-intensive approach to cyber risk identification for AI systems.

---

[101] (Artificial Intelligence Risk & Governance - Artificial Intelligence for Business (upenn.edu)), last visited (22-12-2022).
[102] *See* DNB, *supra* note 10, at 34.
[103] Id, pp. 34-35.

Concerning cyber risk analysis for AI systems, arguments can be made that due to the current state of AI maturity and development in most organisations, analysis of the potential impact of risks should focus on environmental, i.e., external and stakeholder, risk factors rather than more theoretical AI system specific ones since AI systems themselves currently are often not advanced or complex. Still, while AI development and implementation are still very much in development and cyber threats to AI systems may still be mostly theoretical, it is important to be prepared for contingencies. In addition, some researchers find that risk analysis and assessment for AI systems can best be tackled through general risk assessment processes, which allows for leveraging on experience and existing frameworks. Last, regarding risk treatment in which risk mitigation is most relevant, researchers argue that focusing on the creation of an inherently safe design using well known practices from the field of safety studies is an important starting point for cyber risk management for AI-powered systems. Furthermore, various best practices are emerging in cyber risk mitigation for AI-powered systems that should be considered by any organisation (planning on) using AI.

4. Interviews: Cyber risk management for AI systems in practice

The previous chapter examined recent scholarly research on cyber risk management practices relevant to AI-powered systems. Following the three a priori themes identified at the start of the study (technological characteristics, cyber risks, and cyber risk management), and the top-level themes in coding hierarchy of the template analysis, this chapter presents the findings from the five semi-structured interviews, providing insight into the current state of play of cyber risk management for AI-powered systems in the Dutch financial sector.[104] In general, through template analysis, the following were identified in the qualitative interview data:

Themes	Prevalence
Characteristics	
• Data assets	3
• Model assets	3
• AI-powered systems are just another IT asset	3
• Machine learning	2
• Not crown jewels	1
Cyber risk environment	
• Low-risk environments	4
• Model and data risks	3
• Cyber risks from AI receive most attention	2
Cyber risk management	
• Integrated in general (IT) risk management	4
• Risk reduction	2
• Regulatory compliance and license to operate	2
• AI/IT lifecycle management process	2
Challenges in and possibilities for improvement of cyber risk management	
• Diverging and vague regulatory and governance requirements	4
• More regulatory pressure	1
• Additional standards are of little value	1
• Need for more information on cyber risks and their frequency	1

Figure 3: Coding hierarchy.[105]

The figure above shows the top-level themes (underlined) and the lower-level themes grouped and ranked based on shared keywords or key phrases that can trigger a discussion of the (sub-)research question(s). Prevalence refers to the number of interviews in which the themes featured. In the Annex, the themes are clarified through linking them to the qualitative interview data, providing a coding hierarchy. This was then used to create an integrative narrative as laid out in the following paragraphs.

4.1. AI systems used, assets, vulnerabilities, threats, and mitigation

Before presenting the findings on the characteristics of AI systems and their current application in the Dutch financial sector, it is worth noting that all interviewees stated that overall current AI systems used in the sector are neither advanced nor used in what can be considered a high-risk environment.[106] This is illustrated by one expert stating that AI is for PowerPoint and machine learning pipelines are for real life.[107] Not to mention, while all interviewees said to be aware of cyber risks to AI, such as data or model

[104] The interview transcripts are found in Annex 3.
[105] The template analysis process, consisting of first cycle coding, second cycle coding, and the coding hierarchy can be found in Annex 4.
[106] Interview on 10-11-2022; interview on 30-11-2022; interview on 1-12-2022; interview on 2-12-2022; interview on 15-12-2022.
[107] Interview on 10-11-2022.

poisoning attacks, they stated that they have not yet seen such risks and cyber-attacks in the wild.[108] Furthermore, on what kind of AI-powered systems are currently in use in the sector, findings show that in practice most AI can be described as a system incorporating some form of machine learning, with such systems mostly being used in online customer interaction for marketing or help desk purposes, engines for transaction monitoring/categorisation, fraud detection systems, execution only mortgage applications, and cyber security anomaly detection systems.

Concerning the organisational structure of interviewees' financial institutions in relation to cyber risk management for AI systems, all their organisations had a three-line model, with the responsibility for cyber risk management for AI systems being vested in the second line, i.e., risk management and the domain of the Chief Information Security Officer (CISO).[109] While the second line was chiefly responsible for cyber risk management for AI-powered systems in interviewees' organisations, the first line, i.e., the business units, also had responsibilities regarding the operationalisation of cyber risk management, with development teams following security standards and practices such as code scanning for identification of potential security issues as part of the AI development lifecycle.[110] The organisational structure related to cyber risk management for AI systems comes as no surprise since the three-line model is a regulatory requirement for most types of financial institutions. In addition, when it comes to cyber risk management and AI in general, all interviewees noted that most organisations are focused on cyber risks stemming from AI, such as transparency, bias, fairness and explainability, rather than the security of their AI systems and managing cyber risks for AI, indicating that there might be a blind spot.[111]

4.1.1. Characteristics of AI systems

When asked whether their organisation had an AI asset taxonomy such as an overview of specific assets and components used in their AI-powered systems, the practitioners answered that there was no explicit taxonomy or overview in place but that such assets and components were reflected in their management systems or lifecycle management process.[112] In one organisation these assets were simply part of the configuration management database (CMDB), with further risk management tooling like Archer providing assessment, mapping, monitoring, and reporting capabilities for relevant AI asset findings and deviations.[113] Insight into the specific AI assets is gained by them being described in solution designs, part of the lifecycle management system. Concerning which assets their AI systems were most dependent upon, the interviewees provided a broad range of answers, including training data, model artefacts, artefact storage, compute clusters and Kubernetes clusters, application programming interfaces (APIs), stakeholders (i.e. staff who worked on the code) and monitoring processes, with the most material macro-level AI assets being data and model assets.[114]

All interviewees stated that, just like any other IT, their AI systems are subject to a lifecycle management process, tailored to the characteristics of the AI in use. While the lifecycle management processes in place overlapped with many best practices, such as CRISP-DM and MLOps, none of the organisations explicitly based it on one specific standard alone.[115] Furthermore, interviewees stated that the stages of their lifecycle management process are meant to be iterative, with each stage having its own requirements, actions, and approvals. For example, one of the interviewees stated that their AI lifecycle management process starts with an analysis of business needs, followed by drafting requirements,

---

[108] Interview on 2-12-2022.
[109] *See* IIA, *supra* note 67; Interview on 10-11-2022; Interview on 1-12-2022; Interview on 2-12-2022.
[110] Interview on 2-12-2022.
[111] Interview on 10-11-2022; Interview on 2-12-2022.
[112] Interview on 10-11-2022; Interview on 1-12-2022; Interview on 2-12-2022.
[113] (https://www.archerirm.com/), last visited (8-12-2022); Interview on 10-11-2022.
[114] (https://kubernetes.io/), last visited (8-12-2022); Interview on 10-11-2022; Interview on 1-12-2022; Interview on 2-12-2022.
[115] (https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview), last visited (9-12-2022); (https://www.ml4devs.com/articles/mlops-machine-learning-life-cycle/), last visited (9-12-2022); Interview on 10-11-2022; Interview on 2-12-2022.

selection, proof of value, proof of concept, and then a limited implementation of the system, which includes hardening and pen-testing. These steps are followed by learning and training of the model, shadow production to compare functionality with current systems, and finally taking the system into production at which point standard security and cyber risk management controls are applied, such as logging & monitoring, and continuous patching.[116] In general, interviewees stated that each stage of the lifecycle management process had its own IT technical security requirements, such as requiring certain approvals on security scans before a development team could move on to the next stage. Furthermore, in so far possible, the AI lifecycle management process is aligned with project management processes to ensure that governance controls are robust.[117]

4.1.2. Cyber risk for AI systems

Keeping the organisation and its service secure to provide trusted services was stated as the primary concern from a cyber risk point of view.[118] Moreover, cyber risk management for AI is seen as being an inherent part of the sound operational management of financial institutions, with sound operational management being an important regulatory requirement.[119] Within this context, supply chain risk was mentioned as being a focus area, with tracking the integrity of data such as training data, being one of the main concerns regarding the cyber risk for AI systems.[120] Moreover, one of the interviewees stated that cyber risks for AI systems did not receive any specific attention from the development teams, since the platform they worked on has built in controls such as code scanning. As such, cyber risk perception within the business units and their development team is less explicit than in the second line, i.e., security and risk management, since it is part of business-as-usual procedures.[121] Generally, cyber risks for AI systems seem to be considered from a more holistic perspective, with one of the interviewees indicating that cyber risks for AI are receiving increasing attention within the expert community.[122]

Regarding the extent to which cyber risks for AI systems receive specific attention within interviewees' organisations, except for one organisation that considers robust cyber risk management for AI their license to operate, AI currently does not seem to be treated any differently than other IT systems or assets. While interviewees do note that some aspects of cyber risk for AI like model or data risks are incorporated as risk areas in the general IT risk management framework, it is not treated as a field of risk management in itself that requires specific attention such as can be seen with outsourcing risk.[123] One of the interviewees did note that they actively recommend AI model development team members within their organisation to read cyber risk management relevant literature to raise risk awareness in the business units regarding the more ethical and moral dilemmas related to AI.[124]

4.2. Cyber risk management practices

4.2.1. Cyber risk management for AI systems

The interviews find that the main driver behind cyber risk management for AI systems is often risk reduction to ensure business continuity and maintain secure and trusted services, with compliance obligations and public trust also being mentioned. In general, like any other generic risk, cyber risks for AI must fit within the risk appetite of the organisation, in other words, cyber risks for AI must fit within the risk statement of the organisation. Concerning whether senior management prioritised cyber risk management for AI or paid specific attention to it, interview findings show that while the topic does get on the agenda of senior risk committees, the current low maturity of AI system use results in it not

[116] Interview on 10-11-2022.
[117] Interview on 10-11-2022; Interview on 2-12-2022.
[118] Ibid.
[119] Interview on 1-12-2022; interview on 15-12-2022.
[120] Interview on 30-11-2022; Interview on 1-12-2022.
[121] Interview on 2-12-2022.
[122] Interview on 30-11-2022; Interview on 1-12-2022.
[123] Interview on 30-11-2022; Interview on 2-12-2022.
[124] Interview on 10-11-2022.

receiving much senior management attention. Still, recent developments have impacted cyber risk management for AI, with regulatory developments like the proposed AI Act and AI liability Directive likely increasing attention to cyber risks for AI.[125] Moreover, upcoming cyber resilience legislation like DORA seem to have increased overall cyber risk management awareness in the sector.[126]

Concerning how asset vulnerabilities and threats to these vulnerabilities are identified in relation to cyber risk management for AI, interviewees noted that the practices are the same as for all other IT with organisations using, among others, continuous vulnerability scanning tooling that cover their IT landscape, including their AI systems. In general, cyber risk and security management for AI consist of the layered application of traditional risk and security management measures.[127] Any vulnerabilities and threats found are prioritised through standard IT risk management practices, such as vulnerability scoring on CVSS.[128] Interviewees did not mention specific information sources used to gain insight into AI assets, vulnerabilities, and threats, referring to, among others, generic IT risk information repositories or packet & vulnerability management tooling. In general, interviewees stated that the cyber risk management process for AI was not distinct from the process in place for other IT systems or assets.[129] Last, interviewees mentioned that the AI Act will probably lead to some form of AI impact assessment, which could help to identify and prioritise AI asset vulnerabilities and threats.[130]

When asked what frameworks or industry best practices they used, interviewees referred to generic risk management frameworks and standards like those published by NIST, the ISO27001 or ISO27002 standards or the SABSA methodology.[131] One of the interviewees stated that since the AI systems and the environment in which they operate are bespoke, any cyber risk management process or framework must be tailor-made, with the cyber risk management framework consisting of various measures and controls that are mapped on other frameworks. As such, there are very few standards or best practices that can be adopted one to one.[132] When asked whether there were any recent cyber risk management projects that affected their AI systems, one interviewee mentioned that there has been a recent risk management project on fairness and explainability, i.e., risks stemming from AI, that affected the governance framework surrounding their AI systems, establishing more specific risk roles and responsibilities, processes and tooling.[133] This is an example of cyber risks from AI-powered systems receiving attention in lieu of cyber risks for AI systems.

At the end of the interview, interviewees were asked how cyber risk management for AI systems could be improved in their organisation or in the Dutch financial sector. Answers included that there is a need for harmonisation of compliance obligations, preferably on an international level, to create more global uniform requirements and rulings on requirements for AI. Here, reference was made to the General Data Protection Regulation (GDPR) as an influential standard for data protection and privacy. Furthermore, regarding the need for standards and frameworks for cyber risk management for AI, interviewees' opinions varied, with some stating that their organisations or the sector could greatly benefit from AI-specific cyber risk management standards and frameworks, and others stating that additional standards are of little value due to most AI systems and the environments in which they operate being too varied.[134]

---

[125] (https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/), last visited (8-12-2022); (https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en), last visited (9-12-2022); Interview on 10-11-2022; Interview on 30-11-2022; Interview on 2-12-2022.
[126] Interview on 30-11-2022.
[127] Interview on 10-11-2022; Interview 1-12-2022; Interview on 2-12-2022.
[128] (https://www.first.org/cvss/), last visited (9-12-2022); Interview on 10-11-2022.
[129] Interview on 10-11-2022; Interview 1-12-2022; Interview on 2-12-2022.
[130] Interview on 30-11-2022.
[131] (https://sabsa.org/sabsa-executive-summary/), last visited (9-12-2022); Interview on 10-11-2022.
[132] Interview on 2-12-2022.
[133] Ibid.
[134] Interview on 10-11-2022; Interview on 2-12-2022.

One interviewee mentioned that to properly strengthen cyber risk management for AI within their organisation, they would require more regulatory or supervisory pressure so that senior management would allocate budget. In addition, one of the hurdles to better cyber risk management that was mentioned several times was the lack of quantitative data regarding threats to AI systems, which made it difficult to assess which cyber risks are most material and what cyber risk management decisions should be taken based on costs and benefits.[135] Interviewees also stated that more concrete compliance/regulatory requirements on how to deliver certain outputs and/or how to analyse certain things would also be of added value for cyber risk management for AI, with the current regulatory framework being too vague, resulting in adequate or good cyber risk management for AI being too dependent on interpretation. Moreover, more concrete requirements on what can be considered a robust AI-powered system in combination with a certification/assurance scheme was also mentioned as being of added value since this would provide organisations insight into the bounds of cyber risk management for AI.[136]

4.3. Summary

This chapter presented the findings from five interviews with experts and practitioners active in cyber risk management for AI-powered systems in the Dutch financial sector. The interviews found that the use of AI in the sector is still in its early stages, with AI not being considered advanced and predominantly consisting of machine learning algorithms. Most of these systems are implemented in a lower-risk environment that has a human in the loop, such as transaction monitoring and online customer interaction. In most cases, these AI systems are treated like any other IT asset or system, with cyber risk management for AI systems being integrated into the generic IT risk management framework of the organisation. The coverage given to cyber risk management and AI mainly focuses on the management of cyber risks stemming from AI, such as issues pertaining to transparency, bias, fairness and explainability, rather than the management of cyber risk for their systems. This can be explained by the current regulatory framework for AI springing from data protection and privacy law, e.g., the GDPR, which focuses on the impact of technologies like AI on natural persons, AI systems not being advanced and not being used in an organisational high-risk environment, and cyber risks for AI such as data poisoning attacks not being widespread.

Still, interviewees were aware of cyber risks for AI, noting that they do receive (limited) attention within the sector, with data risks and model risks receiving the most consideration. While AI system assets and vulnerabilities are not explicitly tracked on their own, being integrated into more generic asset and vulnerability tracing tooling, they do receive due attention as part of the broader AI lifecycle management process, with these processes often having various built-in security controls and measures, such as code scanning for vulnerabilities or the use of general hardening and pen-testing measures before model deployment. Most cyber risk management controls and measures are based on best practices and industry standards, such as NIST or the ISO27001, with interviewees noting that any risk management process needs to be tailored to the AI system and its environment. Last, interviewees stated that regulatory developments like the AI Act or AI liability Directive result in increased consideration of cyber risk management for AI in the sector, with the expectation that technological developments like more advanced AI, more prevalent use in higher-risk environments and an increase in cyber risks such as data poisoning attacks will result in cyber risk management for AI becoming more important in the next years.

[135] Interview on 2-12-2022.
[136] Interview on 1-12-2022.

5. State of play of cyber risk management for AI systems

The previous chapters presented the findings from the literature study and interviews. This chapter uses these findings to answer the research question and sub-questions, providing insight into gaps and potential areas for improvement in future research and/or cyber risk management practices in the Dutch financial sector. The chapter first examines the theory and interview findings regarding cyber risks for AI systems in the sector, after which the theory and practice of cyber risk management is examined. The chapter ends with a summary of the main findings.

5.1. Cyber risks for AI systems

Answering the first sub-question on which cyber risks AI-powered systems face, the last three years have seen the publication of various academic studies and secondary works on potential cyber risks for AI systems and relevant mitigating measures. When reviewing the literature, this study focused on the AI lifecycle, specific AI assets, their vulnerabilities and potential threats targeting these vulnerabilities. Findings show that not considering any supporting infrastructures such as the protocols enabling system communications, cyber risks for AI systems include data risks, model risks, actor/stakeholder risks, process risks, environment/tool risks, and artefact risks. Moreover, there are also risks for AI emanating from specific ecosystems such as federated learning. Out of these risks, data and model risks receive the most attention, which can be explained by AI systems' dependence on data and model assets, which are also considered to be these systems' defining characteristics. Data and model risks mostly result in the output or outcome changing, impacting an organisations integrity or availability, or in data or model information being leaked, impacting confidentiality. These cyber risks may arise from malicious actors attacking data or model asset vulnerabilities through, e.g., data/model poisoning attacks and/or from non-malicious deficiencies in the lifecycle management process, e.g., inappropriate size definition for stacks or buffers.

Interview findings show that at present, regulatory developments like the AI Act are resulting in increased consideration within the Dutch financial sector of cyber risks for AI systems. Moreover, the cyber risks that receive the most scrutiny are data and model risks, specifically, those data and model risks arising from deficiencies in the lifecycle management process. Cyber risks for AI systems emerging from malicious actors are currently given little notice since cyber-attacks on AI systems or adversarial machine learning is currently not yet observed in practice. This can be explained by most of these systems not being advanced nor being implemented in a high-risk environment where system failure has a material impact on the sound operational management of the organisation. As such, current AI systems are often not considered crown jewels, making these systems an unattractive target for potential attackers. For the sector, this results in a nominal potential organisational impact of cyber risks for AI systems. Still, as AI system complexity advances and their utilisation in high-risk environments increases, cyber risks for these systems will gain more attention. The above results in cyber risks for AI systems in the sector is mostly considered from a conceptual point of view and having little influence on cyber risk management practice.

Cross-comparing findings from the literature study and the interviews, two things stand out, namely: 1) cyber risks originating from malicious actors and their impact on the financial sector currently seem not to be as pertinent as academic publications posit it to be, and 2) cyber risks for AI-powered systems are no different from IT risks in general. First, while the literature mostly approaches cyber risks for AI systems from a cyber-attack/malicious actor point of view, interview findings show that most organisations are focusing on cyber risks for their AI systems arising from negligence or flaws in the lifecycle management process. Moreover, most attention goes towards cyber risks emanating from AI systems and their impact on the public, e.g., issues pertaining to transparency, explainability, bias and fairness. Second, many organisations simply consider AI systems as just another IT asset, and, based on the current maturity level of AI systems in the financial sector, cyber risks for AI systems are often less relevant than cyber risks for other more established IT systems related to, inter alia, cloud

computing systems. As such, any consideration of cyber risks for AI systems should examine relevant risks within the broader context of the environment in which the organisation operates.

5.2. Cyber risk management for AI systems

The answer to the second sub-question, how cyber risks for AI systems can be managed, is that the general conviction in the literature is that cyber risk management for AI systems should be embedded in existing standards and frameworks to leverage on industry experience and to ensure risk management practices are easy to implement. Furthermore, the literature shows that while traditional risk management practices are relevant in managing cyber risks for AI systems, AI systems have unique characteristics, such as its interconnectivity, and reliance on robust data and models, that also need to be considered. This means that in cyber risk management for AI supply chain risks, data and model risks need extra attention, and that a risk management framework lacking mitigating measures and/or controls on these terrains is likely to be inadequate to manage cyber risks for AI-powered systems. To go more in-depth, commonly, cyber risk management can be divided into several risk management steps, of which the steps of risk identification, analysis and treatment are the most relevant for this study. The literature study focused on relevant cyber risk management for AI practices for each of these steps, resulting in several findings.

For cyber risk identification for AI, it can be argued that risk identification should focus on two risk categories inherent to AI systems, namely data-level risk, and model-level risk. This focus allows organisations to tackle the cyber risks stemming from AI systems' unique assets, vulnerabilities, and threats, and allows organisations to increase the trustworthiness and robustness of these systems. Another approach to cyber risk identification for AI is to apply threat modelling to these systems using STRIDE in combination with FMEA to identify cyber risks, analyse, prioritise, and treat them. This provides a structured manner to identify the most material risks, considering their characteristics. Regarding cyber risk analysis for AI systems, findings show that there are roughly two approaches, namely examining likelihood and probability, and focusing on potential impact on the organisation in terms of CIA. For determining likelihood and probability of cyber risk for AI it is possible to use traditional risk analysis practices using a holistic perspective that takes into account environmental/organisational factors and stakeholders involved.

Concerning cyber risk analysis for AI focused on impact, there are multiple tools available such as the use of bug bars for ranking cyber risks for AI. On cyber risk treatment for AI systems, most literature focuses on risk mitigation. Literature shows that for AI systems, a safe design should be the priority, if this is not possible, safeguards should be implemented, and if this is not an option, end users should be informed of relevant cyber risks. Moreover, the safe design of AI systems can be roughly based on four pillars, namely, an inherent safe design, safety reserves, safe failure, and safety-related protective measures. An inherent safe design can be realised through, among others, implementation of relatively uncomplex AI systems, since less complexity allows for easier interpretability and error checking. Safety reserves relate to a built-in tolerance range in mechanical systems, which for AI systems can be translated into ensuring that the uncertainty of the prediction can be measured so that any changes in system performance are spotted. Safe failure relates to ensuring that any failure state of the AI system is mitigated by mechanisms in place such as having a human in the loop where the uncertainty of the prediction is high, or the margin of error is low. Last, safety-related protective measures can be implemented in various manners through, for instance, implementation of a quality assurance process or external protection devices using general security practices.

Interview findings show that in practice, cyber risk management for AI systems is embedded in organisations' general (IT) risk management framework. In most organisations that utilise AI systems, cyber risk aspects for these systems, such as data and model risk, are one of the many risk areas considered as part of managing information risks that an organisation's IT assets are exposed to. The embedment of cyber risk management for AI in the general framework makes sense because IT systems

in general and AI in specific are interconnected. Moreover, cyber risks are often multifaceted, meaning that a holistic approach to cyber risk management is opportune. As such, when answering the third sub-question, what risk management practices do Dutch financial institutions apply to manage cyber risks for AI-powered systems, it is important to consider that the general (IT) risk management practices used for any other information asset also form the foundation for managing cyber risks for AI-powered systems. In addition, the interviews provided three more findings that can answer the third sub-question, namely that: 1) Data and model risk can be a focus area for cyber risk management of AI; 2) robust AI lifecycle management is an essential element of cyber risk management for AI, and 3) that supply chain risk management should receive additional attention within this context.

Findings show that most organisations consider data and model risk to be the most material when managing cyber risks for their AI systems. As a result, in relevant cyber risk management practices, data quality (CIA of data), data validation and model validation receive specific attention. For example, code scanning for vulnerabilities can be an important part of the AI/model lifecycle management process. Furthermore, the interlinkage between cyber risk management for AI and other control management processes should also considered. This is illustrated by interviewees stating that controlling access to AI system data, such as training or test data, is an important practice. Examining AI lifecycle/model management in more detail, interview findings show that a robust lifecycle management process can be central to cyber risk management for AI systems, with robust risk management being dependent on each stage of the lifecycle management process having integrated security controls, such as the abovementioned code scanning but also pre-deployment hardening and pen-testing of AI-powered systems. In general, for robust cyber risk management for AI, the AI/model lifecycle management process should have specific technical security requirements in each stage of the lifecycle. Regarding supply chain risk management being an important aspect in cyber risk management for AI, most organisations make use of third-party service provider products and services as part of their AI system development or implementation. Examples include data being externally sourced, external development of ML models and the use of AI system components like APIs. As such, supply chain vulnerabilities and risks are a key concern when managing cyber risks for AI.

Next to the three findings laid out in the above paragraph, when asked what was necessary for strengthening and/or advancing cyber risk management for AI, interviewees stated that external regulatory and governance changes could be most impactful, with increased regulatory attention being welcomed since it would motivate senior management to better mobilise resources for cyber risk management for AI. Furthermore, interviewees called for more regulatory guidance on the precise requirements for managing cyber risks for AI systems. Since there are many regulatory frameworks, e.g., the GDPR and the proposed AI Act and DORA, with overlapping and distinct legal requirements affecting these practices, practitioners find it difficult to determine which practices fulfil regulatory expectations. In addition, interviewees were divided on the need for more bespoke industry standards or frameworks like NIST, with some welcoming more specific cyber risk management frameworks for AI-powered systems since they could help organisations to strengthen their internal cyber risk management framework, and others stating that cyber risk management and AI are too environment and organisation specific to really benefit from more external standards.

When examining literature and interview findings some differences and gaps become evident. Mainly, AI systems use in terms of the complexity of these systems and the risk environment in which they operate is still in its infancy. As a result, cyber risk management for AI receives little attention, with most cyber risk management practices being integrated in the generic (IT) risk management frameworks of most organisations. Therefore, while interviewees stated that scholarly and technical developments are monitored in the sector, in practice, it is not relevant enough to actively take into consideration in day-to-day cyber risk management. Still, the fact that cyber risk management for AI does not yet play a significant role in practice may also testify to a gap in the cyber resilience of the financial sector. This was also stated by one of the interviewees, who said that while it may not yet be of much relevance, it

is important to keep track of relevant risk developments and actively monitor the situation. That said, it is possible to discern five material findings.

First, both the literature and the practice agree that any cyber risk management practices or framework for AI should be embedded in existing standards and frameworks to leverage the existing risk management environment. Second, literature and practice are consistent that important focus areas in cyber risk management for AI systems are data and model risk since data and model assets are integral to AI system functioning and are a material source of potential risk. Third, next to AI data and model dependencies, its interconnectivity is another important source of risk, resulting in AI supply chain risks being a material area of concern when managing risks for AI systems. Fourth, concerning cyber risk identification and analysis practices, interview findings show that any cyber risk management for AI framework should use an ecosystem perspective that considers the environment in which the system operates, resulting in any cyber risk management framework needing to be bespoke. Fifth, while increased regulatory attention to cyber risk management for AI is generally seen as a good thing, the multifaceted characteristics of AI systems and the risks they face require due consideration of potential contradictory regulatory requirements. Moreover, while technology neutral regulation is important in face of a fast-changing environment, cyber risk management practice can benefit from clear regulatory requirements and/or guidance since vague norms make it difficult to assess when cyber risk management for AI meets expectations.

5.3. Summary

This chapter presented the main findings from the literature study and the interviews, enabling a cross-comparison and providing insight into gaps and potential areas for improvement in future research and/or cyber risk management practice in the financial sector. When cross-comparing literature and practice, what stands out the most is that the maturity of AI use and cyber risk management for AI in the financial sector is low. This is further characterised by AI system complexity in general not being that advanced, its limited implementation within mostly low-risk environments where cyber risk impact on the sound operational management of the organisation is low, and the current perceived risk landscape for AI being conceptual, e.g., few concrete indications of attacks or examined cyber risks for AI-powered systems in practice. However, this does not mean that cyber risk management for AI is an unimportant field of study. All indications point to AI systems becoming more complex and utilisation becoming more widespread. This also means that AI systems will become more interesting for malicious actors, and any cyber risks for AI having a higher potential impact on the organisation, meaning that cyber risk management for AI will also become more important. Furthermore, regulatory developments like the AI Act and DORA make it likely that the field will rapidly develop in the coming two years.

6. Conclusion

6.1. Findings

This study shows that because of AI-powered system assets, vulnerabilities, and threats, there are numerous cyber risks that can impact any organisation using such systems for their business processes. Consideration of cyber risks for AI systems is especially important when the system is complex or used in a high-risk environment, with its functioning having a material impact on the sound operational management of the organisation. As such, appropriate cyber risk management for AI that takes into account relevant cyber risks and AI's unique characteristics, such as their dependence on data and model assets and their possible interconnectedness and complexity, can be of added value to any organisation using such systems. Based on the examined primary academic and secondary sources and following the traditional risk management process of risk identification, analysis, and risk treatment, appropriate cyber risk management for AI is possibly marked by a focus on data-level and model-level risk in the risk identification phase. Moreover, threat modelling using STRIDE and Failure Mode and Effects Analysis can also be of potential added value to risk identification for AI.

For risk analysis for AI, the literature review shows that potential impact on the organisation can likely best be determined by focusing on environmental risk factors and the stakeholders involved since other risk factors such as the technical capacity of AI, are currently not mature. Other potential approaches to risk analysis for AI include the use of bug bars and more established risk analysis processes as can be found in the standards ISO12100 and ISO14971. On risk treatment for AI, most research focuses on risk mitigation techniques rather than risk avoidance, transfer, or risk retention. Based on primary academic and secondary sources, relevant risk mitigation practices for AI include the use of, where possible, AI systems with a low complexity, which better allows for manual error checking, the implementation of safety-related protective measures such as a quality-assurance process, and the use of robust secure software development lifecycle practices or more AI-specific standards like the ISO/IEC TR 24028:2020 for AI system engineering mitigation techniques and methods.

When the results of the literature review are contrasted with the findings from the interviews, several things stand out. Foremost, while public attention to AI, as well as regulatory developments like the AI Act and the Digital Operational Resilience Act, result in cyber risk management for AI-powered systems to become more relevant, the current state of play in the Dutch financial sector is that AI use is not widespread and related cyber risk management practices not being top of mind. Moreover, AI systems in use in the sector are generally not advanced and are implemented in a low-risk environment, e.g., transaction monitoring or chatbots, often having a human in the loop. As such, cyber risks management for AI does not play a significant role in practice. That said, comparing literature and interview findings, it is possible to discern five substantive findings. First, literature and practice agree that any cyber risk management practices or framework for AI should be embedded in extant standards and frameworks to leverage the existing risk management environment.

Second, literature and practice agree that an important focus area in cyber risk management for AI is data and model risk since data and model assets are integral to AI system functioning. Third, next to AI data and model dependencies, AI system interconnectivity is another important source of risk, resulting in AI supply chain risks being a material focus area for cyber risk management. Fourth, concerning cyber risk identification and analysis practices, interview findings show that any cyber risk management for AI framework should use an ecosystem perspective that considers the environment in which the system and organisation operate. Fifth, while increased regulatory attention to cyber risk management for AI is generally seen as a good thing, the multifaceted characteristics of AI systems and the risks they face require due consideration of potential contradictory regulatory requirements. Moreover, cyber risk management practice can benefit from clear regulatory requirements and/or guidance since vague norms make it difficult to assess when cyber risk management for AI meets expectations.

Taking the above findings into account, the brief answer to the main research question 'which cyber risks do AI-powered systems face and how does the Dutch financial sector manage these risks' is that while there are numerous potential cyber risks for AI systems, the low maturity of AI utilisation in the sector and the resultant minimal potential organisational impact of these cyber risks makes cyber risk management for AI mainly a theoretical exercise. Any current cyber risk management for AI practices is firmly embedded in the generic (IT) risk management framework, with most organisations treating AI as just another information IT asset to be considered in their risk management processes. Still, technical, and regulatory developments, such as the AI Act and the Digital Operational Resilience Act, make it likely that cyber risk management for AI-powered systems will gain relevance over the next two years. Nevertheless, whether cyber risk management for AI will become an important risk management area is mostly dependent on organisations utilising it in high-risk environments where it can have a material impact on their sound operational management. Based on the interviews, this is at least another five years away.

6.2. Potential for further research

The literature on cyber risks for AI systems mainly focuses on adversarial machine learning and potential cyber-attacks, such as data and model poisoning attacks. However, few cyber-attacks are yet observed in practice. While it is likely that AI systems will become more interesting targets for malicious actors as their use becomes more widespread and the potential benefits of a successful attack increase, the current practice may benefit more from research on cyber risks for AI emanating from, e.g., deficiencies in system design or environmental factors. In addition, since cyber risk management for AI is generally embedded in the existing (IT) risk management framework of organisations and leveraging on existing standards and practices are seen as important factors in strengthening cyber risk management for AI, research focusing on how to best utilise common industry frameworks and practices for cyber risk management for AI may be of added value rather than focusing on the creation of new frameworks.

Another important area for future consideration, that underlies this study in some respects, is research focusing on how to migrate from a full human-based process to an AI system that is trusted enough to make an autonomous judgement, including which necessary actions are needed to engender this, and how to detect and prevent any unwanted side effects. In other words, what standards do an AI system and any risk management processes surrounding it need to meet from a risk, regulatory and public perspective before it can be implemented in a high-risk environment in which it is trusted enough to make decisions without any human intervention? While widespread adoption of AI systems with the necessary capabilities is some time away, with some experts saying that widespread adoption of such AI is an eternal promise, research focusing on this subject could provide insights of significant value for the field of cyber risk management for AI.[137] In general, the research field is still in its infancy and whatever focus area is chosen there are still many interesting insights to be gained.

[137] (https://fd.nl/tech-en-innovatie/1459071/bruce-sterling-artificial-intelligence-is-de-tulpenmanie-van-nu-orl2cafcPFeO), last visited (18-12-2022).

Bibliography

1. Legislation

Proposal for a 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 (hereinafter referred to as AI Act) (version: compromise text 15 July 2022)

Regulation of the European Parliament and of the Council on digital operational resilience for the financial sector and amending Regulations (EC) No 1060/2009, (EU) No 648/2012, (EU) No 600/2014 and (EU) No 909/2014 (hereinafter referred to as DORA) (version: compromise text 23 June 2022)

2. Books

Russell, S., and Norvig, P., Artificial Intelligence: A Modern Approach, Pearson Higher Education (2021)

Saldaña, J., The Coding Manual for Qualitative Researchers, SAGE 2016

3. Articles

Bouacida, N., and Mohaptra, P., Vulnerabilities in Federated Learning, IEEE Access Volume 9, 2021

Carter, N., Bryant-Lukosius, D., DiCenso, A., Blythe, J., and Neville, A. J., The Use of Triangulation in Qualitative Research, Methods & Meanings, Oncology Nursing Forum (2014)

Datta, S., Shadbolt, N., Backdoors Stuck At The Frontdoor: Multi-Agent Backdoor Attacks That Backfire, arXiv:2201.122211v1

De Silva, D., and Alahakoon, D., An artificial intelligence life cycle: From conception to production, Patterns 3 (2022)

Eling, M., McShane, M., Nguyen, T., Cyber risk management: History and future research directions, Risk Management and Insurance Review (2021)

Elliott, M. W., Risk in an evolving world, The Institutes (2019)

Fredrikson, M., Jha, S., and Ristenpart, T., Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security 2015

Garg, P., Cybersecurity breaches and cash holdings: Spillover effect, Financial Management (2020)

Gordon, L.A., and Loeb, M.P., The Economics of Information Security Investment, ACM Transactions on Information and System Security (2002)

Gu, T., Dolan-Gavitt, B., and Garg, S., BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain, arXiv:1708.06733 (2019)

Haakman, M., Cruz, L., Huijgens, H., and van Deursen, A., AI lifecycle models need to be revised: An exploratory study in Fintech, Empirical Software Engineering (2021)

Harzevili, N.S., Shin, J., Wang, J., and Wang, S., 'Characterizing and Understanding Software Security Vulnerabilities in Machine Learning Libraries', Cornell University, 2022

He, Y., Meng, G., Chen, K., Hu, X., and He, J., 'Towards Security Threats of Deep Learning Systems: A Survey', IEEE Transactions on Software Engineering 2020

Howard, J. D., and Longstaff, T. A., A common language for computer security incidents, Sandia National Labs (1998) (No. SAND98-8667)

Ji, Y., Zhang, Z., Ji, S., Luo, X., and Wang, T., Model-Reuse Attacks on Deep Learning Systems, Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security 2018

Jia, K., and Zhang, N., Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines, Electronic Markets (2022)

Lending, C., Minnick, K., and Schorno, P. J., Corporate governance, social responsibility, and data breaches, Financial Review (2018)

Lin, Q., Verwer, S., Adepu, S., and Mathur, A., TABOR: A Graphical Model-based Approach for Anomaly Detection in Industrial Control Systems, ACM Asia Conference on Computer and Communications Security (2018)

Marotta, A., and McShane, M., Integrating a proactive technique into a holistic cyber risk management approach. Risk Management and Insurance Review (2018)

Mauri, L., and Damiani, E., Modeling Threats to AI-ML Systems Using STRIDE, Sensors (2022)

N. King and J.M. Brooks, Template Analysis for Business and Management Students, SAGE Research Methods 2017

Rosenberg, I., Shabtai, A., Elovici, Y., and Rokach, L., Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain, arXiv:2007.02407 (2021)

Seng Ah Lee, M., and Sing, J., Risk Identification Questionnaire for Detecting Unintended Bias in the Machine Learning Development Lifecycle, Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (2021)

Shokri, R., Stronati, M., Song, C., and Shmatikov, V., Membership Inference Attacks Against Machine Learning Models, In the proceedings of the IEEE Symposium on Security and Privacy 2017

Steimers, A., and Schneider, M., Sources of Risk of AI systems, International Journal of Environmental Research and Public Health (2022)

Strupczewski, G., Defining cyber risk, Safety Science 135 (2021)

Weiss, K., Khoshgoftaar, T.M., and Wang, D., A survey of transfer learning. J Big Data 3, 9 (2016)

Zhang, X., Chan, F.T.S., Yan, C., and Bose, I., Towards risk-aware artificial intelligence and machine learning systems: An overview, Decision Support systems (2022)

4. White papers

Anley, C., Practical Attacks on Machine Learning Systems, NCC Group (2022)

BIS, Working Papers No 865: The drivers of cyber risk (May 2020)

Bundesamt für Sicherheit in der Informationstechnik, Towards Auditable AI Systems: Current status and future directions (2021)

CSR, Adviesrapport Integrale Aanpak Cyberweerbaarheid, 2021

DNB, a macroprudential perspective on cyber risk, Occasional Studies Volume 20 – 1 (2022)

DNB, General principles for the use of Artificial Intelligence in the financial sector (2019)

EBA, EBA Report on Big Data and Advanced Analytics (2020)

ENISA, AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence, December 2020

ENISA, Securing Machine Learning Algorithms, December 2021

ESRB, Mitigating systemic cyber risk, January 2022, pp. 4-25.

National Security Commission on Artificial Intelligence (NSCAI), Final Report, 2021

The AI Index 2022 Annual Report, AI Index Steering Committee, Stanford Institute for Human-Centered AI, Stanford University, March 2022

World Economic Forum, The Global Risks Report 2022: 17th Edition Insight Report

WRR, Opgave AI: De nieuwe systeemtechnologie, 5 November 2021

5. Websites

(Artificial Intelligence Risk & Governance - Artificial Intelligence for Business (upenn.edu)), last visited (22-12-2022)

(https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en), last visited (9-12-2022)

(https://fd.nl/tech-en-innovatie/1459071/bruce-sterling-artificial-intelligence-is-de-tulpenmanie-van-nu-orl2cafcPFeO), last visited (18-12-2022)

(https://kubernetes.io/), last visited (8-12-2022)

(https://learn.microsoft.com/en-us/azure/architecture/data-science-process/overview), last visited (29-10-2022)

(https://learn.microsoft.com/en-us/security/engineering/bug-bar-aiml), last visited (15-10-2022)

(https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml#1-adversarial-perturbation), last visited (15-10-2022)

(https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml?source=recommendations#3-model-inversion-attacks), last visited (16-10-2022)

(https://sabsa.org/sabsa-executive-summary/), last visited (9-12-2022)

(https://www.archerirm.com/), last visited (8-12-2022)

(https://www.bankingsupervision.europa.eu/banking/srep/2021/html/ssm.srep202107_outcomesrepitriskquestionnaire.en.html#toc4), last visited (8-12-2022)

(https://www.consilium.europa.eu/en/press/press-releases/2022/11/28/digital-finance-council-adopts-digital-operational-resilience-act/), last visited (8-12-2022)

(https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/), last visited (8-12-2022)

(https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/), last visited (8-12-2022)

(https://www.dnb.nl/nieuws-voor-de-sector/toezicht-2022/dnb-ziet-cyberdreiging-toenemen-terwijl-basismaatregelen-niet-altijd-op-orde-zijn/), last visited (8-12-2022)

(https://www.first.org/cvss/), last visited (9-12-2022)

(https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview), last visited (29-10-2022)

(https://www.iso.org/iso-31000-risk-management.html), last visited (20-12-2022).

(https://www.iso.org/iso-31000-risk-management.html), last visited (20-12-2022)

(https://www.iso.org/standard/75281.html), last visited (29-10-2022).

(https://www.ml4devs.com/articles/mlops-machine-learning-life-cycle/), last visited (9-12-2022)

(https://www.turing.ac.uk/about-us/frequently-asked-questions), last visited (28-10-2022)

6. Standards and guidelines

IIA, The IIA's Three Lines Model: An update of the Three Lines of Defense (2020)

NPR-ISO/IEC TR 24028, Information technology – Artificial intelligence – Overview of trustworthiness in artificial intelligence (2020)