Towards Model-driven Trustworthiness Assessment of AI Systems

Ijeoma Faustina Ekeh[1,*,†]

[1]Institute of Computer Science, University of Tartu, Estonia

Abstract

This study presents ongoing doctoral research on the security of artificial intelligence systems (hereafter, AI systems). AI security is becoming increasingly important as AI systems are rapidly integrated into critical domains and business operations, demanding robust methods to assess their trustworthiness. While trustworthiness is often considered broadly, it is fundamentally underpinned by security. Recent works lack a systematic way to quantify how well a specific AI system architecture mitigates threats. This research proposed a model-driven framework to assess the security dimension of AI trustworthiness. Using the design science research methodology, the conceptual meta-model and method are developed to guide the evaluation of AI systems. Preliminary results show that security is a foundational pillar of trustworthiness assessment as a pathway to support compliance frameworks.

Keywords

Trustworthiness, AI systems, AI assessment, artificial intelligence security

1. Introduction

The application of artificial intelligence systems has become a key factor in optimising business operations across different tasks. However, this critical integration requires specific steps to provide AI systems with practical robustness against security attacks and other risks. As reported in [1], there is a lack of an adequate basis for assessing AI systems' capabilities.

In the European Union, steps are being taken to regulate the use of these systems, such as the EU AI Act [2], which provides businesses with different risk levels and an enhanced need for vendor assessment protocols [3]. Other organisations have frameworks, such as AI Exchange and MITRE ATLAS [4], primarily for managing security risks and embedding AI security throughout the lifecycle. The NIST AI RMF [5] points to organisations for an implementation of security controls throughout the AI life cycle and provides the ability to incorporate trustworthiness considerations into AI systems. In addition, efforts to implement these frameworks fall short in terms of AI trustworthiness [6].

Frameworks [6, 7] that address all dimensions of trustworthiness, such as security, privacy, fairness, and transparency, provide guidance to industry users of AI systems; however, they introduce overwhelming complexity for industry implementation. Thus, in this research, we focus on the security dimension of trustworthiness to address a specific aspect, namely the domain- and security-specific requirements needed to evaluate the AI system's trustworthiness coverage. By limiting trustworthiness assessment to security, we aim to avoid paralysis when addressing all trustworthiness dimensions, thereby making security a foundational dimension on which other dimensions depend. This research addresses AI system security by proposing a method for assessing their trustworthiness using defined sets of functional and security requirements.

The rest of the paper is structured as follows: Section 2 gives background on AI security and trustworthiness. Sect. 3 describes the method of the PhD research and how the research questions contribute to the research. Sect. 4 presents the ongoing directions of the PhD. Finally, sect. 5 concludes the paper.

Baltic DB&IS 2026 Conference Forum and Doctoral Consortium, 28 June - 1 July 2026, Tartu, Estonia

*Corresponding author.

ijeomafa@ut.ee (I. F. Ekeh)

0009-0003-2145-5856 (I. F. Ekeh)

2. Background

Research on AI security has shifted toward proactive, lifecycle-integrated approaches. Whereas previous AI security efforts focused on protecting trained models and inference endpoints, taxonomic frameworks such as the MITRE ATLAS [4] and governance frameworks such as the NIST AI RMF [5] have recognised that attackers can introduce vulnerabilities at different lifecycle stages of the AI development. Attacks exploiting vulnerable AI system architecture make security a foundational requirement for how businesses integrate them into their operations. Additionally, regulatory drivers such as the EU AI Act's [8] risk-based classification of AI systems and its technical documentation requirements, particularly Annex IV specifications, which mandate evidence of security measures, architectural development, and the structure of AI development, underscore the need to manage AI risks.

The concept of trustworthiness in AI emerges as a multidimensional construct (security, fairness, transparency, reliability, privacy, robustness and safety) [9]. Trustworthiness is the justified confidence [10] that a system performs reliably in alignment with regulatory requirements throughout its lifecycle. For AI to be "trustworthy", it addresses the technical, system-level requirements of an AI system [9, 11]. This enables businesses to incorporate trustworthiness considerations into AI systems. The European Union's High-level Expert Group on AI (HLEG) [10] and NIST AI RMF outline these interdependencies across the different dimensions of trustworthiness, further demonstrating that a deficit in a single dimension can compromise the system's overall reliability. This research scopes AI security as a foundational concept to assessing whether a system could be termed "trustworthy". It explicitly addresses protecting AI system assets, such as architecture and operations, from deliberate harm that could compromise confidentiality, integrity, or availability. Providing a security assessment establishes the extent to which the AI system architecture and operations fulfil defined security requirements against known threats.

Research on AI security and trustworthiness assessment for AI systems has explored different approaches [12, 13, 14, 15]. Kaur et al. [12] examine the need for diverse perspectives on trustworthiness by reviewing trustworthiness requirements and related technical challenges. Similarly, Li et al. [13] present methods to achieve trustworthiness and propose improvements across the AI lifecycle. Different evaluation strategies often capture different perspectives on trustworthiness, e.g., transparency and explainability, ranging from checklists [14] to defined processes [15]. While these existing works provide AI trustworthiness assessment approaches or guidance, they do not offer a method to assess the degree to which an AI system's identity architecture fulfils the security requirements needed to ensure security and robustness. Our work bridges this gap by first presenting a model that captures the core architecture of an AI system, linking assets, lifecycle states, and the operations performed within each lifecycle. Using the defined architecture, our work formalises the requirements – functional and security – to identify gaps in the AI system's trustworthiness. Unlike prior work, security in our context is treated as a needed architectural property that can be measured to ensure compliance and improved iteratively.

3. Research approach

The PhD research follows the Design Science Research Methodology (DSRM) [16] (see Fig.1), which provides a structured framework for artefact development and research conduct. The method addresses practical and theoretically grounded problems in artificial intelligence security and emphasises the purpose of the developed artefacts in delivering solutions to the identified issues. Here, the development of the AI trustworthiness assessment framework follows the key stages of DSRM, including a systematic literature review, method evaluations, and testing using real-world LLM-based systems and deployments in test environments and industry.

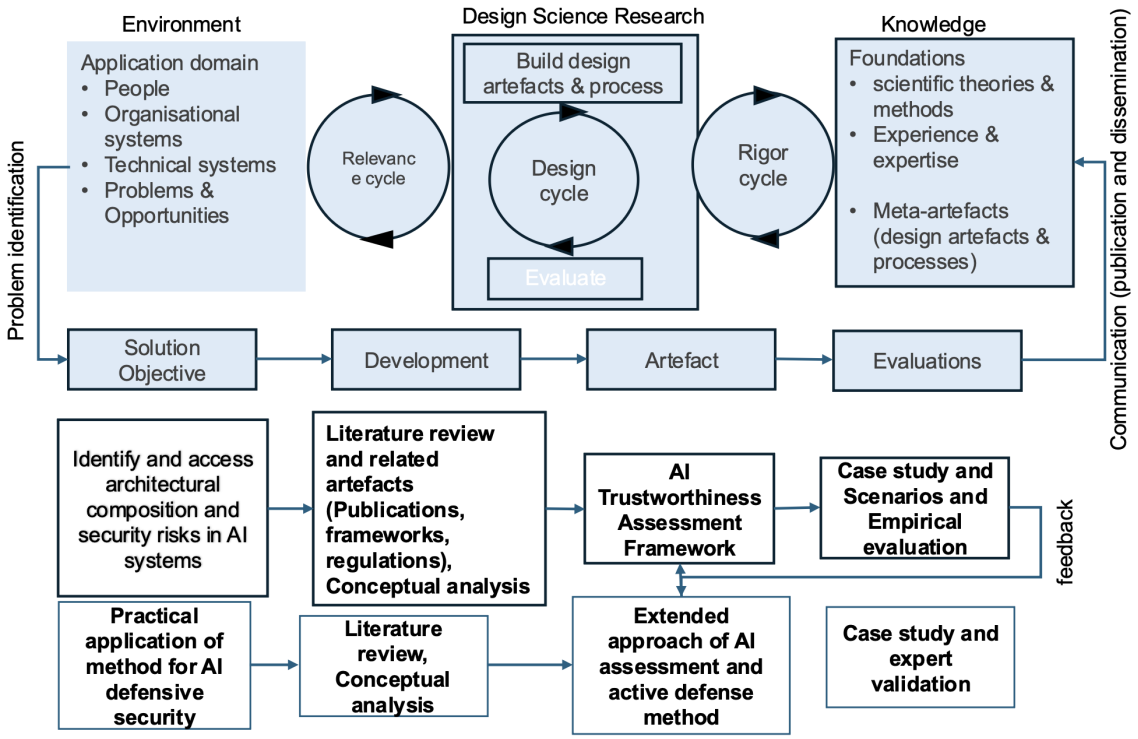

Figure 1: PhD Research Structure [16].

3.1. Problem identification

The current gap in AI security approaches, where organisations cannot determine the extent to which their AI systems fulfil the security requirements necessary to be considered trustworthy, poses the following questions:

3.1.1. Research questions

Four questions guide the PhD research. Each question provides a sequence for developing and testing the AI trustworthiness assessment framework.

RQ1: How can we systematically assess the coverage of functional and security requirements in AI systems to determine their trustworthiness? This question intends to investigate and address how we can bridge the gap in how businesses assess AI before integrating it into their business operations. The activities that precede this phase include a systematic literature review that provides the context and state of the art in AI security, which, in turn, leads to the conceptual development of an AI system asset model [17] and the assessment framework.

RQ2: How can the model-driven trustworthiness assessment framework be applied through an AI-powered tool, and how do the coverage scores of this automation enable specific governance mechanisms? This question provides further development of the established artefacts in my first contribution. This phase investigates how large language models can be constrained within a meta-model to conduct reliable, structured security trustworthiness assessments and to proffer a solution that enhances the integration of governance frameworks, thereby increasing the robustness of the assessment framework. The validation of the work would compare automated systems with manual expert assessments across diverse AI systems.

RQ3: How can governance frameworks be effectively integrated into the assessment of AI security and trustworthiness? The activities of RQ2 provide an essential bridge to making the framework scalable and feasible in governance. My research aims to ensure the assessment framework's consistency and scalability by integrating with and aligning with regulations such as the EU AI Act and NIST RMF, enabling seamless adaptation.

RQ4: What ways could the trustworthiness coverage drive real-time AI defensive measures against evolving threats in the AI system? This question addresses the goal of closing the loop between assessing an AI system's trustworthiness and the active defence of the system based on identified gaps. For example, a low coverage score for model poisoning controls would trigger adaptive responses, such as temporarily disabling the model or isolating the affected API endpoint where the attack could occur.

3.2. Solution objective

The primary objective of this research is to develop an AI trustworthiness Assessment framework, a conceptual framework aimed at providing businesses with a comprehensive assessment of the architectural, security risks, and mitigation coverage of AI systems, and to guide them towards actionable governance objectives.

3.3. Design and development

The design and development phase of the AI trustworthiness assessment framework is based on a systematic review of the literature, conceptual analysis, and background studies of AI security and trustworthiness frameworks, guidelines, and regulations. The activities developed three artefacts: the conceptual meta-model used for the instantiation of the AI systems; requirement elicitation and threat mappings; and, finally, the method for qualitative evaluation of trustworthiness coverage. Furthermore, to enable adoption, an automated tool is needed to interpret AI system documentation and conduct system evaluation. To further extend the development, the study aims to develop a conformance framework that enhances the scalability of the AI trustworthiness assessment for AI security governance. The evolution of the assessment tool is aimed at practical defensive actions to enable a model-informed active defence, in which security controls adapt based on function, and was conducted to identify gaps in security risk management for AI systems.

The Conceptual Meta-Model Figure 2 presents the proposed AI system asset model [17], which serves as the basis for assessing the trustworthiness of AI systems. The meta-model, represented as a UML class diagram, defines the core functions and operations of an AI system, including training and inference. The model is depicted as a UML diagram that describes the static architecture of AI system assets, operations, and their relationships.

System assets represent the technical components of an AI system. These include: The machine learning (ML) model as a component that supports inference; the ML processing system, which facilitates the inference and other supporting IT infrastructures (e.g., GPUs); the machine learning training system, which performs the training of the ML model; and the ML system input API, which covers the APIs necessary for inference and user inputs. The business assets, such as training data, input data, intellectual property, and embeddings, are supported by the system assets. This meta-model provides a framework for instantiating an AI system's architecture using its publicly available documentation.

3.4. Demonstration and evaluation

To establish a foundation for the trustworthiness assessment framework, a systematic literature review [17] following Kitchenham's methodology was conducted to identify gaps in security risk management for AI systems. This process led to the conceptual development of the meta-model, structured around

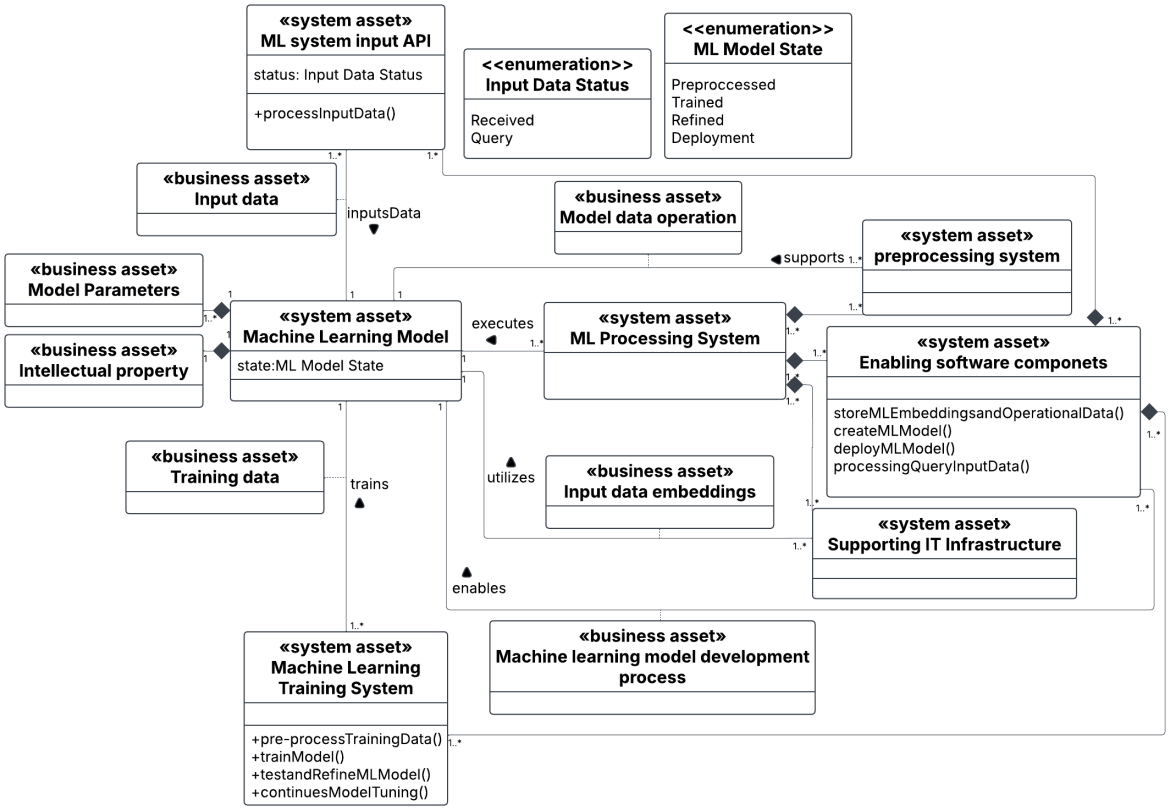

Figure 2: AI System Asset Meta-model.

the Information Systems Security Risk Management (ISSRM) [18] and the various AI lifecycle processes, depicted as 'operations'. The defined concepts of the meta-model led to the elicitation of functional and security requirements used to evaluate AI systems and determine their trustworthiness coverage. Following the development of the conceptual model as a driver for the assessment, a proposed method to assess the trustworthiness coverage of the AI system is based on defined coverage metrics (see Fig. 3). The phases include:

- **Instantiation of the meta-model:** The available documentation provided by these systems is classified using the architecture (covering the system and business assets and the operations the system undergoes) of AI systems into the established meta-model.
- **AS-IS Evaluation:** Evaluation of the system based on the defined functional and security requirements to establish the coverage of the system.
- **Security Analysis:** Conduct security analysis on the AS-IS state of the AI system to identify the security threats posed to the system and possible mitigations.
- **TO-BE Evaluation:** Evaluation using the functional and security requirements on the applied security controls for the AI system.
- **Trustworthiness Coverage (TC):** This is defined as the weighted combination score of the functional coverage and security coverage.

$$\text{TC} = (FC, SC) \tag{1}$$

Where: FC and SC represent the functional and security coverage of requirements for the accessed AI system's architecture across both AS-IS and TO-BE evaluations.

To demonstrate the method's use, eight LLM-based systems, including open-weight models such as Mistral AI and Llama 3.3, were evaluated. The AI system instantiations are based on available documentation; an AS-IS evaluation against functional and security requirements [17] using the conceptual

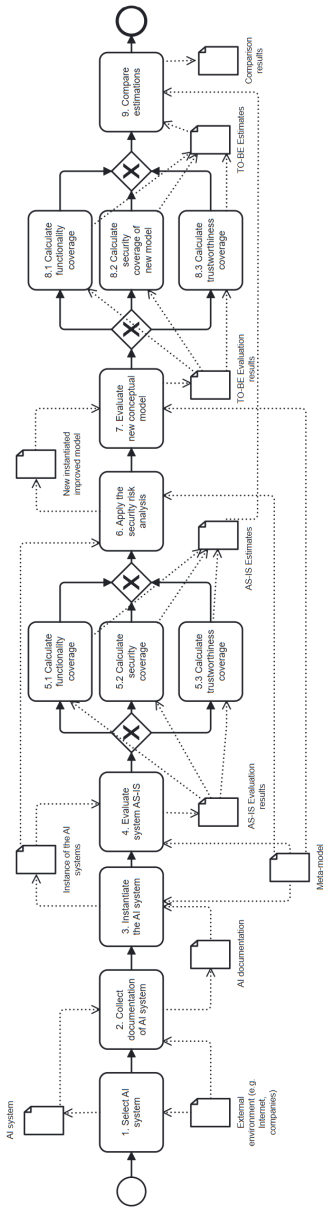

Figure 3: Method for Model-driven AI Trustworthiness Coverage Evaluation.

meta-model; a security risk analysis; and a TO-BE evaluation with applied security controls. Figure 4 presents the results of the 8 assessed LLM-based systems – Mistral AI, Baseline T5, Duplex, Llama-3.2, Llama-3.3, OpenELM, Pythia, and Qwen3. To support the comparative assessment across the eight evaluated LLM-based systems, we represent trustworthiness as a two-dimensional construct combining functional coverage (FC) and security coverage (SC). Each dimension is segmented into three fulfilment bands (0–39.99%, 40–60.99%, 61–100%), and the intersection of these bands produces the composite colour regions.

The AS-IS results in Figure 4 demonstrate the relative maturity of functional coverage in LLM-based systems, with several achieving above 60%. Mistral AI, Duplex, and Pythia exhibited the most consistent functionality coverage in the high functionality band (61–100%), whereas Llama-3.2, Llama-2.2, Baseline T5, and Qwen-3 fall into the medium band (40–60.99%). However, OpenELM demonstrated the lowest coverage of 38.2%, confirming that the documented architecture configurations lacked comprehensive safeguards. In contrast, the security assessment shows a markedly different pattern. Every system is positioned within the lowest security band (0–39.99%), meaning that none of the architectures meet partial security requirements in their AS-IS state. The AS-IS evaluation shows clear

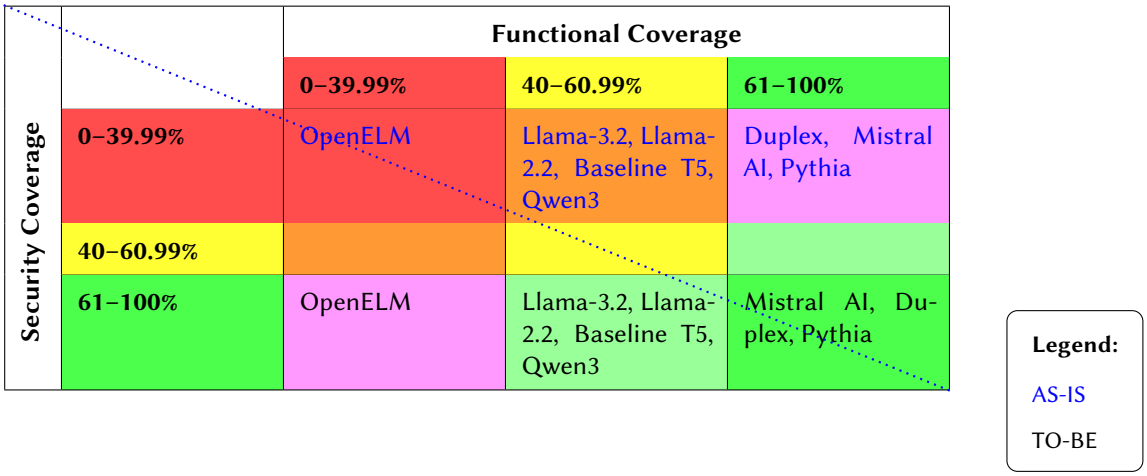

Figure 4: Assessment of Coverage of Functionality and Security Requirements in LLM-Based systems.

variation in functional maturity but uniformly insufficient security coverage, resulting in an imbalance in trustworthiness across all systems. After applying the proposed method, the TO-BE results show substantial improvement in the security dimension (see Table 4). All systems transition from the lowest security band (0–39.99%) to the highest band (61–100%) while their functional coverage bands remain unchanged. These shifts yield distinct changes in the composite regions. This presents systems with high functionality, such as Mistral AI, Duplex, and Pythia, now occupying green composite cells (high FC × high SC), indicating balanced fulfilment of functional and security requirements.

4. Work in progress

The immediate work in the research focuses on two areas to address RQ2 and RQ3: designing and implementing an automated trustworthiness assessment process to define a pathway to operational impact for businesses. Concurrently, the research seeks to determine how the trustworthiness coverage outputs directly inform and activate governance mechanisms. To scale the assessment to integrate security-focused governance, regulations such as the EU AI Act [19, 2] explicitly address the security of AI systems; the plan is to align the AI trustworthiness assessment framework as a benchmark for scaling security due diligence and regulatory compliance for AI systems. The research would explore how to integrate trustworthiness assessments into an actionable governance framework.

Further research will leverage automated assessment to explore how they could drive active defence through the configurations of monitoring systems and inform adaptive responses to protect AI systems.

5. Conclusion

This PhD research presents ongoing research on a model-driven framework for assessing AI trustworthiness. The work proposes a method for evaluating the trustworthiness coverage of AI systems (in the case of the preliminary evaluation, LLM-based systems) using defined sets of functional and security requirements. Immediate research work includes automating the method as an integral part of integrating governance mechanisms and enhancing its use as an active defence tool for AI systems. The research proposes to address the identified gap in translating descriptive AI security frameworks and regulations into actionable, qualitative security assurance guidance relevant to threats and mitigation for industry needs. The framework provides businesses with a structured methodology that explicitly models system components and operations, derives quantifiable requirements, and evaluates their fulfilment through the specified coverage metrics. Future work will refine and extend the framework through continuous validation in industry AI systems and explore its scalability for AI governance.

Acknowledgments

This PhD thesis is supervised by Prof. Raimundas Matulevičius at the Institute of Computer Science, University of Tartu, Estonia. The European Union funds this research under Grant Agreement No. 101087529. However, views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible.

Declaration on Generative AI

The author have not employed any Generative AI tools.

References

[1] A. Omran Almagrabi, R. A. Khan, Optimizing Secure AI Lifecycle Model Management With Innovative Generative AI Strategies, IEEE Access 13 (2025) 12889–12920. URL: https://ieeexplore.ieee.org/document/10742321. doi:10.1109/ACCESS.2024.3491373.

[2] Section 2: Requirements for High-Risk AI Systems | EU Artificial Intelligence Act, 2024. URL: https://artificialintelligenceact.eu/section/3-2/.

[3] C. M. Pierson, E. Hildt, From Principles to Practice: Comparative Analysis of European and United States Ethical AI Frameworks for Assessment and Methodological Application, Proceedings of the Association for Information Science and Technology 60 (2023) 327–337. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pra2.792. doi:10.1002/pra2.792, _eprint: https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/pra2.792.

[4] MITRE ATLAS™, 2025. URL: https://atlas.mitre.org/.

[5] NIST Risk Management Framework Aims to Improve Trustworthiness of Artificial Intelligence, NIST (2023). URL: https://www.nist.gov/news-events/news/2023/01/nist-risk-management-framework-aims-improve-trustworthiness-artificial, last Modified: 2025-02-03T17:30-05:00.

[6] C. Frischknecht-Gruber, P. Denzel, M. Reif, Y. Billeter, S. Brunner, O. Forster, F.-P. Schilling, J. Weng, R. Chavarriaga, AI Assessment in Practice: Implementing a Certification Scheme for AI Trustworthiness (Academic Track), OASIcs, Volume 126, SAIA 2024 126 (2025) 15:1–15:18. URL: https://drops.dagstuhl.de/entities/document/10.4230/OASIcs.SAIA.2024.15. doi:10.4230/OASICS.SAIA.2024.15, artwork Size: 18 pages, 2257072 bytes ISBN: 9783959773577 Medium: application/pdf.

[7] D. Korobenko, A. Nikiforova, R. Sharma, Towards a Privacy and Security-Aware Framework for Ethical AI: Guiding the Development and Assessment of AI Systems, in: Proceedings of the 25th Annual International Conference on Digital Government Research, dg.o '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 740–753. URL: https://dl.acm.org/doi/10.1145/3657054.3657141. doi:10.1145/3657054.3657141.

[8] High-level summary of the AI Act | EU Artificial Intelligence Act, 2024. URL: https://artificialintelligenceact.eu/high-level-summary/.

[9] N. Kemmerzell, A. Schreiner, H. Khalid, M. Schalk, L. Bordoli, Towards a Better Understanding of Evaluating Trustworthiness in AI Systems, ACM Comput. Surv. 57 (2025) 218:1–218:38. URL: https://dl.acm.org/doi/10.1145/3721976. doi:10.1145/3721976.

[10] Ethics guidelines for trustworthy AI | Shaping Europe's digital future, 2018. URL: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[11] P. Vassilakopoulou, E. Parmiggiani, A. Shollo, M. Grisot, Responsible AI: Concepts, critical perspectives and an Information Systems research agenda, Scandinavian Journal of Information Systems 34 (2022). URL: https://aisel.aisnet.org/sjis/vol34/iss2/3.

[12] D. Kaur, S. Uslu, K. J. Rittichier, A. Durresi, Trustworthy Artificial Intelligence: A Review, ACM Comput. Surv. 55 (2022) 39:1–39:38. URL: https://dl.acm.org/doi/10.1145/3491209. doi:10.1145/3491209.

[13] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy AI: From Principles to Practices, ACM Comput. Surv. 55 (2023) 177:1–177:46. URL: https://dl.acm.org/doi/10.1145/3555803. doi:10.1145/3555803.

[14] Lumenalta, AI security checklist (updated 2025) | Strengthen defenses with a solid AI security checklist | Secure your processes with AI security checklist LLM | Adopt an AI security & governance checklist LLM approach, 2025. URL: https://lumenalta.com/insights/ai-security-checklist-updated-2025, section: Insights.

[15] R. V. Zicari, J. Brodersen, J. Brusseau, B. Düdder, T. Eichhorn, T. Ivanov, G. Kararigas, P. Kringen, M. McCullough, F. Möslein, N. Mushtaq, G. Roig, N. Stürtz, K. Tolle, J. J. Tithi, I. van Halem, M. Westerlund, Z-Inspection®: A Process to Assess Trustworthy AI, IEEE Transactions on Technology and Society 2 (2021) 83–97. URL: https://ieeexplore.ieee.org/abstract/document/9380498. doi:10.1109/TTS.2021.3066209.

[16] K. Peffers, T. Tuunanen, M. A. Rothenberger, S. Chatterjee, A Design Science Research Methodology for Information Systems Research, Journal of Management Information Systems 24 (2007) 45–77. URL: https://doi.org/10.2753/MIS0742-1222240302. doi:10.2753/MIS0742-1222240302, _eprint: https://doi.org/10.2753/MIS0742-1222240302.

[17] I. F. Ekeh, R. Matulevicius, Systematic Technical Report for Model-driven Elicitation of Functional and Security Requirements in AI-supported Systems, Technical Report, Zenodo, 2025. URL: https://zenodo.org/records/17878604. doi:10.5281/zenodo.17878604.

[18] N. Mayer, Model-based Management of Information System Security Risk, 2012. URL: https://www.semanticscholar.org/paper/Model-based-Management-of-Information-System-Risk-Mayer/66c31937f0a8734842ff115bc0e38d90ff2b833a.

[19] Annex IV: Technical Documentation Referred to in Article 11(1) | EU Artificial Intelligence Act, 2024. URL: https://artificialintelligenceact.eu/annex/4/.