Proceedings of the 35th European Safety and Reliability & the 33rd Society for Risk Analysis Europe Conference
Edited by Eirik Bjorheim Abrahamsen, Terje Aven, Frederic Bouder, Roger Flage, Marja Ylönen
©2025 ESREL SRA-E 2025 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-94-3281-3_ESREL-SRA-E2025-P7802-cd

AI Security Assurance: Developing Framework for Secure and Resilient AI

Ankur Shukla

Department of Risk and Security, Institute for Energy Technology, Norway. E-mail: ankur.shukla@ife.no

Shao-Fang Wen, and Basel Katt

Department of Information Security and Communication Technology, Norwegian University of Science and Technology, Norway. E-mail: shao-fang.wen@ntnu.no, basel.katt@ntnu.no

The rapid advancement of Artificial Intelligence technologies has delivered considerable transformative benefits across various industries but has also brought significant security risks. Security assurance of AI systems is critical, particularly as these systems are increasingly integrated into critical infrastructures, healthcare, financial services, and autonomous systems. This paper discusses the challenges, risks, and opportunities related to AI systems, covering various aspects such as data preprocessing, model training, and deployment. It presents a conceptual framework for AI security assurance, focusing on evaluating the overall security level based on security requirements, threats, vulnerabilities, and ethical considerations. The framework leverages established security standards, regulations, and acts to identify security requirements and provide a structured approach to identifying and addressing AI-specific risks. The paper aims to provide insights into security risks related to AI and highlight the importance of incorporating security assurance measures throughout the AI system lifecycle.

Keywords: Artificial Intelligence, AI Security Assurance, Security Risks, Trustworthiness, Risk Management.

1. Introduction

Artificial Intelligence (AI) systems are rapidly transforming industries and societies, driving advancements in automation, decision-making, and predictive analytics. However, as the adoption of AI systems grows, so do the associated risks and vulnerabilities. AI systems are susceptible to adversarial attacks, data poisoning, model inversion, and evasion tactics that can compromise their security, integrity, and reliability (Goodfellow et al., 2014). Therefore, it is imperative to establish a comprehensive framework to ensure the security and resilience of AI systems.

Some efforts have been made to ensure AI assurance and trustworthiness. Stettinger et al. (2024) proposed methods to ensure the trustworthiness of high-risk AI systems, ensuring compliance with the EU's AI Act by fulfilling trustworthiness requirements throughout their lifecycle. Similarly, Gadewadikar et al. (2023) introduced an AI assurance method that considers five components: ethics, transparency, compliance, safety, and certification. However, current approaches often focus on isolated aspects, such as adversarial robustness, model explainability, or data integrity, without offering an integrated, end-to-end approach to securing AI systems (Hernández-Rivas et al., 2024; Papernot et al., 2016). Traditional security methodologies, such as threat modeling, risk assessment, and testing protocols, are not fully equipped to address the unique attack surfaces and dynamic nature of AI systems (Huang et al., 2011). Additionally, the complexity of AI models, coupled with the lack of transparency and interpretability, makes it difficult to detect, prevent, and mitigate potential attacks (Doshi-Velez and Kim, 2017). As AI systems become more autonomous, ensuring their security and resilience is essential for maintaining public trust and regulatory compliance.

This paper aims to address the security assurance challenges of AI systems by developing a comprehensive security assurance framework for security and resilient AI. It discusses the various AI security risks and ethical considerations and presents a conceptual framework that integrates security, threats, and ethical considerations throughout the AI lifecycle. The paper also aligns the proposed framework with the established framework of AI, such as the NIST AI RMF. The framework guides developers, architects, and security experts in building secure and resilient AI systems.

2. AI Systems and Security Assurance

AI systems are computational systems designed to simulate human cognitive abilities, such as learning, reasoning, and problem-solving (Martínez-Fernández et al., 2022). These systems leverage algorithms, statistical models, and large datasets to achieve autonomy in decision-making and predictive analytics (Manickam et al., 2022; Russell and Norvig, 2016). AI systems are typically classified into narrow AI, which performs specific tasks like image recognition or natural language processing, and general AI, which aspires to have the reasoning capabilities of a human being (Nilsson, 2009). Modern AI systems rely on a combination of machine learning, deep learning, and reinforcement learning techniques to adapt and improve their performance over time (Goodfellow, 2016). These systems have found applications in diverse fields, such as healthcare, finance, transportation, and cybersecurity, demonstrating their potential to enhance productivity and innovation.

The architecture of AI systems typically involves multiple components, including data ingestion, data preprocessing, feature extraction, model training, and model inference (Chollet and Chollet, 2021). The models used in AI systems can be categorized into supervised, unsupervised, and reinforcement learning models, each designed to address different types of problems (Bishop and Nasrabadi, 2006). The complexity of these systems introduces several challenges in terms of interpretability, security, and robustness, which have become key focus areas in AI system development. It is important to develop comprehensive frameworks for their security assurance to address issues related to adversarial attacks, data poisoning, and privacy concerns (Oseni et al., 2021; Papernot et al., 2016).

2.1. *Defining AI System Security Assurance*

We defined the AI system security assurance as follows: "*The confidence that an AI system meets its security and ethical requirements, effectively mitigates risks, and maintains resilience against vulnerabilities, threats, and ethical violation throughout its lifecycle.*"

2.2. *Risk, Challenges, and Opportunities in AI System Security*

The deployment of AI systems introduces a wide spectrum of risks, challenges, and opportunities in system security. One significant risk is the susceptibility of AI systems to adversarial attacks, where small perturbations to input data can lead to incorrect predictions or classifications (Chen et al., 2019; Goodfellow et al., 2014). Such attacks can severely impact AI-driven applications in healthcare, finance, and autonomous vehicles, where incorrect outputs may have severe consequences. For example, adversarial attacks on image recognition systems involve adding subtle perturbations to images such that the AI model misclassifies them. For example, a subtle change in a stop sign image might cause the model to classify it as a yield sign (Wang et al., 2024). Moreover, AI systems are vulnerable to data poisoning (Acuña, 2024), where adversaries manipulate training data to embed malicious behaviors into models. These risks highlight the critical need for robust security mechanisms, especially in safety-critical applications.

One of the key challenges in AI system security is ensuring interpretability and explainability. Due to the "black-box" nature of many AI models, especially deep learning models, it is difficult to trace how specific predictions are made (Lipton, 2018). This opacity limits the ability of developers and auditors to detect and mitigate attacks or errors. Moreover, the dynamic learning capability of AI models adds complexity to security assurance. AI models constantly evolve and improve, which makes it challenging to maintain a consistent security level over time. Adversarial attacks on these models also vary depending upon the type of algorithm used. In essence, while threats to AI systems are generally similar, the methods of exploiting them differ according to the specific algorithm being used (Oseni et al., 2021).

Despite these risks and challenges, AI systems also present opportunities for enhanced security measures. For instance, AI-driven threat detection systems can identify and mitigate attacks in real-time using anomaly detection and behavioral analytics (Olabanji et al., 2024). AI-enabled security tools can predict potential vulnerabilities before exploitation, allowing organizations to adopt a proactive security posture (Manoharan and Sarker, 2023). Furthermore, explainable AI (XAI) research aims to bridge the interpretability gap, enabling transparent and justifiable AI decision-making processes (Arrieta et al., 2020). By incorporating explainability, organizations can not only detect attacks more effectively but also comply with regulatory requirements for accountability and transparency. These opportunities highlight the potential of AI not only as a system to be secured but also as a transformative enabler of security advancements.

3. Proposed Framework for AI Security Assurance

This section covers the proposed AI security assurance framework (Fig. 1), based on our earlier framework for cyber-physical and IT systems (Shukla et al., 2023; Wen et al., 2022; Katt and Prasher, 2019).

3.1. *Security Assurance Program*

The Assurance Program provides a structured framework to ensure the security, compliance, and ethical integrity of AI systems. It includes the policies, procedures, and tools to methodically evaluate, manage, and improve security measures, with an emphasis on trust, compliance, and risk management. It aligns with established security frameworks and complies with key standards and regulations.

3.2. *Security Assurance Profile*

The Security Assurance Profile (Katt and Prasher, 2019) outlines the AI system's security context, objectives, and requirements, providing a clear and comprehensive understanding of its security posture. The key components of the security assurance profile are:

3.2.1. *System Overview*

The system overview defines the AI system's purpose, function, and alignment with business goals, including its technical architecture, AI models, data sources, and integrations. It also maps data flows, detailing inputs, processing, outputs, and storage, and specifies the operational environment (on-premises, cloud, or hybrid)along with external dependencies and interactions (Manickam et al., 2022).

3.2.2. *Security Objectives*

The security assurance profile defines the required security objectives that ensure the AI system's security, resilience, and trustworthiness. These objectives include confidentiality, integrity, availability, authentication and authorization, and accountability.

3.2.3. *Security Assurance Requirements*

The security assurance profile defines key security requirements for AI systems, including confidentiality, integrity, availability, authentication, authorization, and accountability. It also incorporates the requirements and best practices and complies with industry standards and regulations like ISO 42001, ISO 27001, the EU AI Act, and GDPR, ensuring the system's security, resilience, and ethical compliance.

3.2.4. *Ethical Requirements*

The security assurance profile also includes the key ethical requirements to ensure that the AI system operates in a responsible and transparent manner (Wen et al., 2025).

- ***Privacy***: To ensuring that the AI system handles sensitive data in compliance with privacy laws and industry best practices.
- ***Bias***: To address potential biases in AI model training and decision-making while ensuring fair and equitable results for diverse populations.
- ***Transparency***: To ensure the explainability and interpretability of AI models, it is essential that

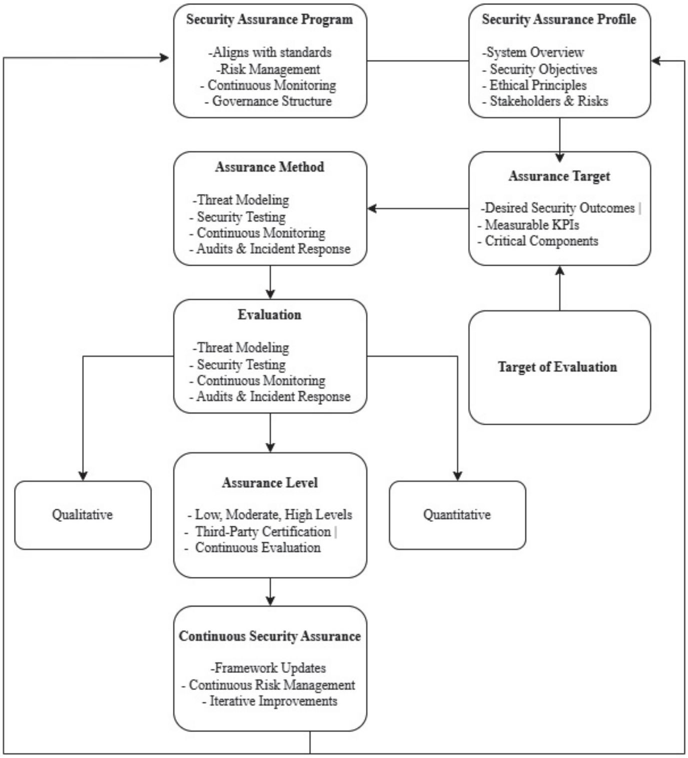

Fig. 1. Proposed framework for AI security assurance.

stakeholders can understand and trust the system's decisions and predictions.

3.2.5. *Expectations from the Stakeholders*

The security assurance profile considers the requirements and expectations of key stakeholders, including users, operators, and regulatory bodies. These expectations may include data privacy, system reliability, ethical AI use, robust system management and security by operators, compliance with legal and industry standards, and proactive measures to address potential threats from adversaries (Kinney et al., 2024).

3.2.6. *Risk Assessment*

Risk assessment identifies and mitigates potential threats and vulnerabilities against AI systems. This process is helpful to identify the necessary security requirements to ensure comprehensive risk management and achieve the expected level of system security and resiliency (Khlaaf, 2023; Wen et al., 2025).

3.3. *Security Assurance Metrics*

Security assurance metrics (Katt and Prasher, 2019) are important in evaluating the level of assurance of an AI system's security and ethical compliance. These metrics are used to measure and assess the system's performance and resilience. Some of the security assurance metrics are as follows:

- ***Security Requirements Metrics***: This metric is calculated based on the identified requirements that the AI system must meet.
- ***Ethical Requirements Metrics***: This metric is calculated based on the identified ethical requirements.
- ***Threat and Vulnerability Metrics***: This metric is calculated based on various threat analysis and vulnerability testing techniques, including threat modeling and analysis, vulnerability scanning,

and security testing.

The relationships among security assurance requirements, threats, vulnerabilities, and ethical considerations are interdependent and mutually reinforcing. Security assurance requirements are designed to mitigate identified threats and vulnerabilities, ensuring that the AI system remains resilient against potential attacks. These security measures are implemented with ethical considerations in mind, as ethical principles guide the development and application of security practices. This ensures privacy, transparency, and protection against bias. Moreover, ethical considerations influence how threats and vulnerabilities are addressed, especially when they affect privacy or fairness.

3.4. *Security Assurance Target*

The assurance target (Katt and Prasher, 2019) defines the desired security outcomes and objectives for the AI system, specifying what must be assured to ensure its effectiveness and resilience. This includes measurable security goals like model robustness, data integrity, adversarial resilience, and confidentiality, along with expectations for handling adversarial inputs, securing critical components, and ensuring secure external interactions.

3.5. *Target of Evaluation*

The Target of Evaluation (ToE) defines the specific AI system components, processes, and functionalities subjected to security assessment. It identifies the boundaries of what is being evaluated to ensure the system's security, resilience, and compliance with defined assurance goals. The ToE includes AI models, data pipelines, interfaces/APIs, the operational environment, and controls to protect against threats and breaches (Xia et al., 2024).

3.6. *Security Assurance Methods*

The security assurance methods (Katt and Prasher, 2019) outline the techniques, tools, and practices used to evaluate and validate the security of AI systems. These are some of the methods that can be used to assess the security assurance level of AI systems:

3.6.1. *Security Requirements Verification*

Security requirements verification is an important method to ensure the security of an AI system. It involves verifying that the system meets predefined security requirements, identifying requirements that are not met, and identifying potential vulnerabilities.

3.6.2. *Threat Modeling & Risk Assessment*

This involves identifying potential threats, attack vectors, categorizing risks based on their potential impact and likelihood using different techniques such as STRIDE, DREAD, OCTAVE, FAIR, etc Atmaca et al. (2022); Katt and Prasher (2019); Shukla et al. (2022).

3.6.3. *Security Testing*

Security testing is an essential method for the security assurance of AI systems. This includes penetration testing, vulnerability scanning, adversarial testing, and other techniques to evaluate and ensure the system's security and resilience Shukla et al. (2022); Wen et al. (2025).

3.6.4. *Audits & Compliance Checks*

Audits and compliance checks are the key methods and play a critical role in the security assurance of the AI systems. Audits includes the expert evaluations of security policies, controls, and system architecture that ensure the alignment with best practices and security standards. On the other hand, the compliance checks confirm that the AI system meets the relevant security standards and regulatory requirements including ISO 27001, ISO 42001, GDPR, and the EU AI Act (Falco et al., 2021; Lam et al., 2024).

3.7. *Security Assurance Evaluation*

Security assurance evaluation (Katt and Prasher, 2019) is the process of assessing and evaluating the AI system's security posture based on the evidence collected using security assurance methods. It utilizes two primary evaluation methods: quantitative and qualitative.

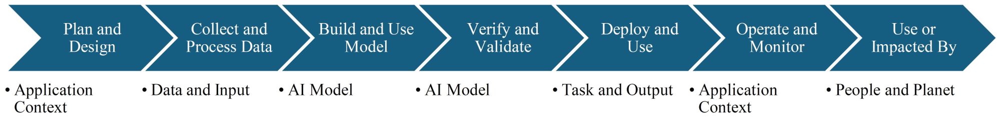

Fig. 2. AI systems lifecycles and key dimensions.

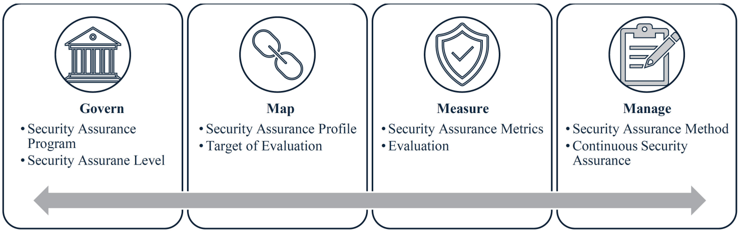
Govern
• Security Assurance Program
• Security Assurane Level
Map
• Security Assurance Profile
• Target of Evaluation
Measure
• Security Assurance Metrics
• Evaluation
Manage
• Security Assurance Method
• Continuous Security Assurance

Fig. 3. Mapping NIST AI RMF and the proposed security assurance framework.

3.8. *Assurance Level*

The assurance level (Katt and Prasher, 2019) describes the degree of confidence in the AI system's ability to meet its security objectives based on the results of the assessment. The confidence levels range from basic assurance, if significant risks have not been addressed, to high assurance, when strong security controls have been implemented with low residual risks. Third-party certification and validation are also used to provide external verification of the AI system's security posture.

3.9. *AI System Continuous Security Assurance*

Continuous Security Assurance is a dynamic, ongoing process that ensures AI systems remain secure as they evolve Shukla et al. (2022); AI (2023); Wen et al. (2025). It involves different activities such as

- **Regularly update** the security framework to address emerging threats, regulatory changes, and technological advancements.
- **Continuous risk management** to identify, assess, and mitigate the risks to maintain system resilience, and
- **Iterative improvements** to refine security practices based on feedback from testing, audits, and responses to security incidents.

4. AI Security Assurance and NIST AI Framework

The NIST AI RMF (AI, 2023) is a voluntary framework designed to integrate trustworthiness into the lifecycle of AI systems. It defines key dimensions and lifecycles (Fig. 2) and is built around four core functions: govern, map, measure, and manage. The proposed Security Assurance Framework aligns closely with the NIST AI RMF to ensure secure and ethical AI development (Fig. 3).

- ***Govern:*** The Govern function establishes policies, processes, and accountability to manage AI System risks. The Security Assurance Program can be aligned with this function by providing a structured framework with policies, tools, and procedures to ensure security, compliance, and ethical integrity, emphasizing trust and risk management. Furthermore, Assurance Levels set benchmarks through confidence levels and third-party certifications.
- ***Map:*** The Map function identifies and contextualizes AI risks. The Security Assurance Profile can be aligned with this function by defining the AI system's security context, objectives, and requirements, offering a comprehensive view of its security posture, including risk assessments. On the other hand, ToE maps specific system components

and boundaries for assessment.

- ***Measure:*** The Measure function assesses and tracks AI risks. Security Assurance Metrics and Evaluation can be aligned with this function by providing measurable indicators of an AI system's security and ethical compliance, covering performance, resilience, and vulnerabilities.
- ***Manage:*** The Manage function mitigates AI risks based on impact. The AI System Security Assurance Methods and Continuous Security Assurance can be aligned with this function by detailing techniques like threat modeling, security testing, audits, and ongoing risk management to ensure resilience and security.

5. Conclusion and Future Work

This paper presents a conceptual framework for AI security assurance, addressing challenges, risks, and ethical considerations in securing AI systems. The framework provides a structured approach to identifying and mitigating AI-specific security threats by leveraging established security standards, regulations, and best practices. Furthermore, it aligns with the NIST AI RMF to enhance deployment and ensure mutual complementarity.

However, as the framework is currently conceptual, its real-world applicability remains to be validated. Future work will focus on empirical validation through case studies and real-world implementations. Conducting experimental evaluations across diverse AI-driven domains such as healthcare, finance, and autonomous systems will provide practical insights into the framework's effectiveness and adaptability. Additionally, collaboration with industry partners to test the framework in live AI environments can help refine its components and ensure practical relevance. By incorporating empirical data and real-world case studies, the framework can be further developed into a robust security assurance model that bridges the gap between theoretical security considerations and practical AI deployment challenges.

References

Acuña, E. G. A. (2024). Healthcare cybersecurity: Data poisoning in the age of ai. *Journal of Comprehensive Business Administration Research*.

AI, N. (2023). Artificial intelligence risk management framework (ai rmf 1.0).

Arrieta, A. B., N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion 58*, 82–115.

Atmaca, U. I., C. Maple, G. Epiphaniou, et al. (2022). Challenges in threat modelling of new space systems: A teleoperation use-case. *Advances in Space Research 70*(8), 2208–2226.

Bishop, C. M. and N. M. Nasrabadi (2006). *Pattern recognition and machine learning*, Volume 4. Springer.

Chen, T., J. Liu, Y. Xiang, W. Niu, E. Tong, and Z. Han (2019). Adversarial attack and defense in reinforcement learning-from ai security view. *Cybersecurity 2*, 1–22.

Chollet, F. and F. Chollet (2021). *Deep learning with Python*. simon and schuster.

Doshi-Velez, F. and B. Kim (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Falco, G., B. Shneiderman, J. Badger, R. Carrier, A. Dahbura, D. Danks, M. Eling, A. Goodloe, J. Gupta, C. Hart, et al. (2021). Governing ai safety through independent audits. *Nature Machine Intelligence 3*(7), 566–571.

Gadewadikar, J., J. Marshall, Z. Bilodeau, and Vatatmaja (2023). Systems engineering–driven ai assurance and trustworthiness. In *Conference on Systems Engineering Research*, pp. 343–356. Springer.

Goodfellow, I. (2016). Deep learning.

Goodfellow, I. J., J. Shlens, and C. Szegedy (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Hernández-Rivas, A., V. Morales-Rocha, and J. P. Sánchez-Solís (2024). Towards autonomous cybersecurity: A comparative analysis of agnostic and hybrid ai approaches for advanced persistent threat detection. In *Innovative Applications of Artificial Neural Networks to Data Analytics and Signal Processing*, pp. 181–219. Springer.

Huang, L., A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. D. Tygar (2011). Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, pp. 43–58.

Katt, B. and N. Prasher (2019). Quantitative security assurance. In *Exploring security in software architecture and design*, pp. 15–46. IGI Global.

Khlaaf, H. (2023). Toward comprehensive risk assessments and assurance of ai-based systems. *Trail of Bits 7*.

Kinney, M., M. Anastasiadou, M. Naranjo-Zolotov, and V. Santos (2024). Expectation management in ai: A framework for understanding stakeholder trust and acceptance of artificial intelligence systems. *Heliyon 10*(7).

Lam, K., B. Lange, B. Blili-Hamelin, J. Davidovic, S. Brown, and A. Hasan (2024). A framework for assurance audits of algorithmic systems. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1078–1092.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue 16*(3), 31–57.

Manickam, P., S. A. Mariappan, S. M. Murugesan, S. Hansda, A. Kaushik, R. Shinde, and S. Thipperudraswamy (2022). Artificial intelligence (ai) and internet of medical things (iomt) assisted biomedical systems for intelligent healthcare. *Biosensors 12*(8), 562.

Manoharan, A. and M. Sarker (2023). Revolutionizing cybersecurity: Unleashing the power of artificial intelligence and machine learning for next-generation threat detection. *DOI: https://www. doi. org/10.56726/IRJMETS32644 1*.

Martínez-Fernández, S., J. Bogner, X. Franch, M. Oriol, J. Siebert, A. Trendowicz, A. M. Vollmer, and S. Wagner (2022). Software engineering for ai-based systems: a survey. *ACM Transactions on Software Engineering and Methodology (TOSEM) 31*(2), 1–59.

Nilsson, N. J. (2009). *The quest for artificial intelligence*. Cambridge University Press.

Olabanji, S. O., Y. Marquis, C. S. Adigwe, S. A. Ajayi, T. O. Oladoyinbo, and O. O. Olaniyi (2024). Ai-driven cloud security: Examining the impact of user behavior analysis on threat detection. *Asian Journal of Research in Computer Science 17*(3), 57–74.

Oseni, A., N. Moustafa, H. Janicke, P. Liu, Z. Tari, and A. Vasilakos (2021). Security and privacy for artificial intelligence: Opportunities and challenges. *arXiv preprint arXiv:2102.04661*.

Papernot, N., P. McDaniel, A. Sinha, and M. Wellman (2016). Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*.

Russell, S. J. and P. Norvig (2016). *Artificial intelligence: a modern approach*. Pearson.

Shukla, A., B. Katt, L. O. Nweke, P. K. Yeng, and G. K. Weldehawaryat (2022). System security assurance: A systematic literature review. *Computer Science Review 45*, 100496.

Shukla, A., B. Katt, and M. M. Yamin (2023). A quantitative framework for security assurance evaluation and selection of cloud services: a case study. *International Journal of Information Security 22*(6), 1621–1650.

Stettinger, G., P. Weissensteiner, and S. Khastgir (2024). Trustworthiness assurance assessment for high-risk ai-based systems. *IEEE Access 12*, 22718–22745.

Wang, T., Z. Bi, Y. Zhang, M. Liu, W. Hsieh, P. Feng, L. K. Yan, Y. Wen, B. Peng, J. Liu, et al. (2024). Deep learning model security: Threats and defenses. *arXiv preprint arXiv:2412.08969*.

Wen, S.-F., A. Shukla, and B. Katt (2022). Developing security assurance metrics to support quantitative security assurance evaluation. *Journal of Cybersecurity and Privacy 2*(3), 587–605.

Wen, S.-F., A. Shukla, and B. Katt (2025). Artificial intelligence for system security assurance: A systematic literature review. *International Journal of Information Security 24*(1), 1–42.

Xia, B., Q. Lu, L. Zhu, and Z. Xing (2024). An ai system evaluation framework for advancing ai safety: Terminology, taxonomy, lifecycle mapping. In *Proceedings of the 1st ACM International Conference on AI-Powered Software*, pp. 74–78.