Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 682–692
https://doi.org/10.54941/ahfe1005635

Optimizing AI System Security: An Ecosystem Recommendation to Socio-Technical Risk Management

Kitty Kioskli[1], Antonios Ramfos[2], Steve Taylor[3], Leandros Maglaras[4], and Ricardo Lugo[5]

[1]trustilio B.V., Amsterdam, The Netherlands
[2]Athens Technology Center, Athens, Greece
[3]School of Electronics and Computer Science, University of Southampton, UK
[4]School of Computing Edinburgh Napier University Edinburgh, UK
[5]Estonian Maritime Academy, Tallinn University of Technology, Tallinn, Estonia

ABSTRACT

Given the sophistication of adversarial machine learning (ML) attacks on Artificial Intelligence (AI) systems, enhanced security frameworks that integrate human factors into risk assessments are critical. This paper presents a comprehensive methodology combining cybersecurity, cyberpsychology, and AI to address human-related aspects of these attacks. It introduces an AI system security optimization ecosystem to help security officers protect AI systems against various attacks, including poisoning, evasion, extraction, and inference. The risk management approach enhances NIST and ENISA frameworks by incorporating socio-technical aspects of adversarial ML threats. By creating digital clones and using explainable AI (XAI) techniques, the human elements of attackers are integrated into security risk management. An innovative conversational agent is proposed to include defenders' perspectives, advancing the design and deployment of secure AI systems and guiding future certification schemes.

Keywords: AI system security, Socio-technical risk management, Explainable AI (XAI), Cybersecurity frameworks

INTRODUCTION

Artificial Intelligence (AI) systems, which typically process data to generate predictions, recommendations, or decisions based on statistical models, are becoming indispensable across sectors like commerce, healthcare, and defense. These systems rely on machine learning (ML) models trained on extensive datasets to perform tasks that traditionally required human intelligence. However, their growing prevalence has also made them prime targets for adversarial attacks. Adversarial ML attacks exploit vulnerabilities in ML models and training datasets, aiming to manipulate the system's outputs. These attacks can occur at any stage of the ML lifecycle, from training to deployment. During training, adversaries might introduce malicious data to corrupt the model—a tactic known as poisoning attacks. During deployment, attackers can craft inputs specifically designed to

mislead the model into making incorrect predictions or decisions, known as evasion attacks (Demetrio et al., 2021). Human error and oversight often exacerbate these vulnerabilities, as they are frequently exploited to facilitate such breaches.

Evidence from real-world scenarios suggests that many organizations, despite being aware of adversarial ML threats, lack the necessary tools and methodologies to protect their AI systems effectively. Papernot et al. (2016) demonstrated that even well-established AI systems are vulnerable to relatively simple adversarial techniques, highlighting an urgent need for more sophisticated defensive measures. Additionally, Biggio and Roli (2018) point out that the absence of standardized security practices across organizations further exacerbates the vulnerability of AI systems. To address these challenges, there is an urgent need for affordable and effective cybersecurity practices that can enhance the resilience and security of AI systems against adversarial attacks. This paper proposes a novel approach that integrates human factors into the security risk assessment process. By combining insights from cybersecurity, cyberpsychology, and AI, we aim to develop a comprehensive methodology to tackle the human-related aspects of adversarial ML attacks.

This paper presents an AI system security optimization ecosystem to help security officers protect AI systems from adversarial attacks, including poisoning, evasion, extraction, and inference. It extends NIST and ENISA frameworks by incorporating socio-technical dimensions. By using digital clones and ML models trained on data from cyber exercises, it integrates human elements of attackers into security risk management. Explainable AI (XAI) techniques are proposed to monitor vulnerabilities, increasing transparency and trust. Additionally, an innovative conversational agent will include the human perspective of defenders, promoting a holistic approach to AI security. The aim is to advance the design and deployment of secure AI systems and guide the development of future certification schemes, promoting the broader adoption of AI technologies in a secure and trustworthy manner. Our research envisions to not only contribute to the technical robustness of AI systems but also enhance the socio-technical framework necessary for their safe integration into society. This comprehensive approach is crucial to ensuring that AI can achieve its full potential while mitigating the risks associated with adversarial threats.

OBJECTIVES & AIMS

This paper aims to comprehensively manage the technical and human-related aspects of adversarial ML threats by leveraging the AI Risk Management Framework proposed by NIST (NIST_AI_RMF), along with information security risk management and AI standards such as ISO31000, the ISO2700x family, ISO/IEC WD 27090, ISO/IEC WD 27091, ISO/IEC 22989, ISO/IEC FDIS 5338 AI, and CEN/CLC/JTC 21 AI Risk Management, and the Multilayer AI Framework from ENISA. The primary objective is to propose cost-effective defense methods that enable AI stakeholders and security

officers to effectively safeguard AI systems against both technical and socio-technical threats arising from evolving adversarial ML attacks, whether during development, deployment, or active attack scenarios.

The ecosystem this paper proposes includes methodologies designed to: Efficiently Elicit Defense Socio-Technical Needs; Conduct Affordable Socio-Technical Adversarial ML Attack Analysis; Enhance AI Security Risk Management and Continuously Evaluate Security Risks.

INTEGRATED SOCIO-TECHNICAL RISK EVALUATION FOR AI SYSTEM SECURITY

The proposed ecosystem, illustrated in Figure 1, empowers security officers to significantly enhance the security of AI systems through continuous, iterative cycles of risk assessment. These cycles integrate essential human factors and incrementally apply appropriate mitigation measures until a satisfactory level of security is achieved. The socio-technical AI system risk assessment evaluates the risk level of potential attacks and recommends actions to mitigate these risks by considering both the severity and likelihood of such attacks. This assessment is informed by the attackers' socio-technical profiles and the vulnerabilities identified in the AI system using advanced XAI techniques.

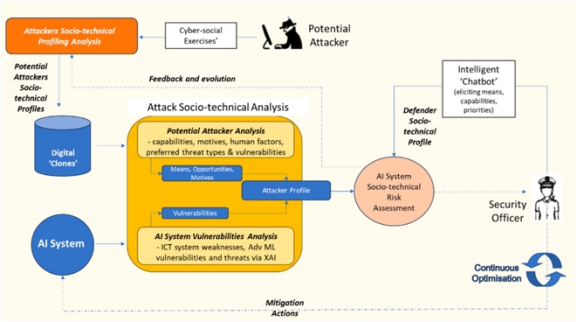

Figure 1: Conceptual architecture.

A critical aspect of this paper is overcoming the intrinsic difficulty in identifying and integrating the socio-technical profiles of attackers. To address this, digital clones provide insights into the means, motives, and opportunities of potential adversarial ML attackers. Additionally, the proposed socio-technical AI system risk assessment incorporates the socio-technical profile of defenders by using an intelligent chatbot to interact with security officers. This interaction enables the dynamic updating of attacker profiles based on feedback from ongoing socio-technical risk assessments regarding the likelihood and nature of potential attacks. Implementing this concept requires addressing several key challenges: Identification of Relevant Human Aspects of Adversarial ML Attacks; Incorporation of Socio-Technical Aspects into the Risk Management Process of AI Systems and Experimentation-Based Approach to AI System Security Optimization.

By addressing these challenges, the proposed ecosystem aims to enhance the security and resilience of AI systems against adversarial ML attacks comprehensively. The continuous risk assessment cycles, informed by socio-technical profiles and supported by advanced XAI techniques, provide a robust framework for identifying and mitigating vulnerabilities. This ecosystem ensures that AI systems can operate securely in diverse environments, maintaining an optimal balance between security measures and system performance. The integration of human factors into the risk management process represents a significant advancement in the field, promoting the development of more secure and trustworthy AI technologies.

IDENTIFYING HUMAN FACTORS IN ADVERSARIAL ML ATTACKS: PSYCHOSOCIAL AND BEHAVIORAL ANALYSIS OF ATTACKERS AND DEFENDERS

Methodologies from investigative psychology and behavioural sciences provide valuable tools for understanding the behaviours and motivations of adversarial machine learning (ML) attackers and defenders. Sanders and Stappers (2019) suggest using profiling techniques and social experiments to gather insights into the psychosocial characteristics of individuals involved in adversarial ML attacks. These methodologies involve structured interviews, behavioural observations, and psychological assessments to identify the cognitive and emotional factors that influence adversarial behaviours. By leveraging these insights, researchers can develop more accurate profiles of attackers and defenders, informing the development of targeted defensive strategies.

Before enhancing AI system risk management to include valuable human aspects, it is crucial to address the challenge of identifying these aspects during an adversarial ML attack. The sophistication and potentially prolonged duration of such attacks make detection and monitoring difficult, complicating the identification and integration of credible information about the psychosocial and behavioural characteristics of adversarial ML attackers and the security operators defending against these attacks (Demetrio et al., 2021).

To address this problem of adversarial ML attacks and better understand the 'enemy,' this framework proposes the following innovative approach: Literature and Specification Review; Enhanced Profiling through Cyber-Social Exercises; Knowledge Generation and Data Analysis and Model Training.

This paper also suggests the incorporation of investigative psychology research and behavioural science for psychosocial and behavioural analysis, utilizing accurate behavioural models like Fogg's Behavioural Model. Fogg's model posits that the likelihood of a behaviour (B) occurring is a product of Motivation (M), Ability (A), and the appropriate Trigger (T) (Fogg, 2019). Based on this model, human profiles can be developed using five categories of traits: Personality, Social-Behavioural, Technical Awareness, Motivation, and Trigger, each with specific attributes and measurement scales. Therefore,

this paper suggests the set-up of cyber-social exercises involving face-to-face interviews with potential attackers and defenders to build these profiles.

Existing research on adversarial ML attacks underscores the importance of human-related aspects in understanding and mitigating these threats. Pierazzi et al. (2019) discuss the need for robust adversarial defences that account for human decision-making processes and perceptions. Their formalization of problem-space attacks highlights the relationship between feature space and problem space, providing a foundation for more principled research in this domain. Similarly, Dyrmishi et al. (2023) emphasize the importance of human perceptibility as a criterion for assessing the effectiveness of adversarial attacks, underscoring the role of human factors in evaluating and improving model robustness. Additionally, the psychosocial and behavioural characteristics of adversarial ML attackers are crucial for developing effective defensive strategies. Attackers often exhibit high technical proficiency; driven by a deep understanding of the ML algorithms and systems they seek to compromise. Their motivation can stem from various sources, including financial gain, personal satisfaction, or ideological beliefs. Papernot et al. (2016) highlights the importance of understanding these motivations to anticipate and counter adversarial actions. Moreover, specific behavioural triggers can prompt adversarial activities, such as perceived vulnerabilities in the system or opportunities for significant impact. These factors must be considered in defensive strategies to effectively anticipate and mitigate attacks. Additionally, understanding the cognitive processes that underlie adversarial behaviours, such as decision-making under uncertainty and risk perception, can provide valuable insights for developing more robust AI defences. Studies Demetrio et al. (2021) emphasize the need for comprehensive profiling of adversarial actors to inform defensive measures and enhance system resilience.

Incorporating human aspects into AI system risk management is essential for enhancing the robustness and reliability of these systems. Effective risk management frameworks must consider human-related aspects to address the complexities of adversarial ML attacks. Standards and frameworks such as ISO 31000 for risk management, ISO/IEC 27001 for information security management, and the NIST AI Risk Management Framework emphasize the importance of integrating human factors into the risk management process (ISO, 2018; NIST, 2021). These frameworks provide guidelines on managing risks faced by organizations, highlighting the need to consider human behaviours and decision-making processes. Incorporating human factors involves understanding the psychosocial dynamics that influence adversarial behaviours, such as motivation, cognitive biases, and social interactions. By leveraging methodologies from investigative psychology and behavioural sciences, organizations can develop more comprehensive risk management strategies that account for the complex interplay between human and technological factors. For instance, Sanders and Stappers (2019) suggest using behavioural profiling techniques to gather insights into the motivations and behaviours of adversarial ML attackers and defenders, enabling more effective risk mitigation strategies. Data from cyber-social exercises can be used to train ML models that replicate the characteristics

of adversarial ML attackers, improving the ability to predict and mitigate attacks. These models can simulate the decision-making processes and behavioural patterns of adversarial actors, providing valuable insights for developing targeted defence strategies. Additionally, 'question-and-response' schemes can effectively elicit human aspects from defenders, enhancing their ability to anticipate and counter adversarial tactics (Papernot et al., 2016).

Having the competencies to utilize data from cyber-social attack exercises and a comprehensive understanding of the NIST AI Risk Management Framework, along with the understanding of human elements (i.e., biases, profiles), may also help mitigate the recent technological development focusing on deceptive language models (DLLMs) that use backdoor attacks. These backdoor attacks employ covert, natural, and highly invisible triggers that blend seamlessly with normal data, making them challenging to detect (Hubinger et al., 2024). Techniques like homograph replacement or using subtle differences between text generated by language models and natural text create triggers that are visually indistinguishable from legitimate content. These triggers achieve high success rates with minimal data injection, complicating standard data inspection techniques and bypassing collaborative AI and human detection models. Advanced methods such as steganography and regularization create invisible backdoors, embedding triggers in ways that circumvent human perception and standard detection tools. Dynamic triggers that adapt to specific contexts or inputs further complicate detection. Additionally, backdoor attacks are designed to maintain high invisibility scores, ensuring that triggers do not stand out even during close human inspection (Hubinger et al., 2024). These strategies exploit the misalignment between machine learning models' sensitivities and human perceptual abilities, making it extremely difficult for human administrators to detect them (Li et al., 2021; Sun et al., 2022; Yang et al., 2021).

EXPLAINABLE AI FOR CYBERSECURITY (XAI)

The key elements of cybersecurity modeling include automation, which minimizes manual efforts through self-learning; intelligence, which supports informed decision-making based on extracted insights; and trustworthiness, which ensures human-interpretable cyber decisions. These aspects enable efficient and effective protection against evolving threats in increasingly complex digital environments. Therefore, balancing "Automation," "Intelligence," and "Trustworthiness" is crucial. A more transparent and understandable AI model, known as Explainable AI (XAI), can enhance the effectiveness of cybersecurity modeling. Analysts and security professionals can utilize this information to understand system operations, identify potential vulnerabilities and threats, and make optimal actionable decisions. This section examines AI and XAI-based methods for cybersecurity modeling and their potential real-world applications, considering the key aspects of XAI for cybersecurity.

To comprehend the potential of diverse AI methods, we first classify them into six key categories based on their working principles, as outlined

in (Sarker, 2024). These categories are Machine Learning, Deep Learning, Large Language Models (LLMs), Rule-Based Systems, Semantic Knowledge Representation, and Uncertainty Modeling. Each category has its pros and cons, as identified in (Sarker, 2024). Recent studies in cybersecurity indicate a growing interest in leveraging graph structures to improve the detection and recognition of network attacks (Gilliard, 2024).

By encoding domain-specific knowledge and enabling intelligent decision-making, semantic knowledge representation and reasoning provide a strong basis for advancing cybersecurity modeling. Semantic technologies, including ontologies (formal representations of knowledge within a specific domain) and knowledge graphs (structured representations that capture entities, their attributes, and their relationships), facilitate rich data integration and analysis. These technologies enable cybersecurity models to capture intricate relationships among threats, vulnerabilities, assets, and defensive measures.

Semantic techniques are applied in various areas such as security monitoring and malware analysis, enhancing the capability to address complex cybersecurity challenges. This provides sophisticated reasoning capabilities, enabling models to infer complex insights, identify potential attack scenarios, and recommend tailored countermeasures based on contextual understanding. However, designing efficient algorithms and scalable inference mechanisms is crucial for detecting anomalies, identifying patterns, and inferring actionable insights from large-scale semantic knowledge bases. Machine learning and knowledge or rule mining methods can further augment knowledge graphs through tasks such as entity linking, node classification, relation extraction, recommendation, searching, disambiguation, feature engineering, and construction automation. These advancements make such applications more useful and effective for cybersecurity-oriented applications. The application of XAI will help our system perform better in detecting early complex attack scenarios and present the data in a understandable manner to the security officers.

DEFENDERS' ANALYSIS

To incorporate the human aspects of defenders into the security risk assessment of AI systems, this paper proposes the use of advanced conversational agents or 'chatbots' to facilitate intelligent dialogues. These dialogues aim to achieve two main objectives: firstly, to elicit the security operator's capabilities in defending the ML system, and secondly, to gather valuable information about the business environment's priorities where the ML system is deployed.

Recent advancements in natural language processing (NLP) have led to the development of a new generation of conversational agents, or intelligent chatbots, which are significantly more adept at understanding natural language and generating more engaging and contextually appropriate responses. This progress has made intelligent chatbots increasingly popular across various domains, including business, healthcare, and education. Intelligent chatbots can be broadly categorized based on their goals and technical approaches. Goal-wise, they are divided into task-oriented chatbots,

which are designed to accomplish specific tasks, and conversational or open-ended chatbots, which are intended for more general and flexible interactions. Technically, chatbots are classified into three types: rule-based, retrieval-based, and generative-based.

Empirical evaluations suggest that task-oriented chatbots perform best when implemented with rule-based approaches, as these ensure reliability and predictability in executing specific tasks. Conversely, conversational chatbots benefit from generative AI-based approaches, which allow for more dynamic and natural interactions. By leveraging these advanced conversational agents, the paper aims to enhance the security risk assessment process. The intelligent chatbots will interact with security operators to uncover their defensive capabilities and understand the business priorities that influence the security landscape. This dual approach ensures a comprehensive integration of human factors into the risk assessment, leading to more robust and context-aware security strategies for AI systems. Overall, the integration of intelligent chatbots into security risk assessments represents a significant innovation, capitalizing on the latest advancements in NLP and AI to address the complex and dynamic nature of cybersecurity threats.

SOCIO-TECHNICAL RISK ASSESSMENT TOOLING AND KNOWLEDGE ENGINEERING

The methodology outlined in this paper is designed to manage adversarial ML threats by integrating several key frameworks and standards. These include the AI Risk Management Framework proposed by NIST (NIST_AI_RMF) and information security risk management standards such as ISO 31000, ISO 27001, and ISO 27005 (see Figure 2). Additionally, the methodology incorporates specific standards used throughout the lifecycle of AI systems as defined by ENISA.

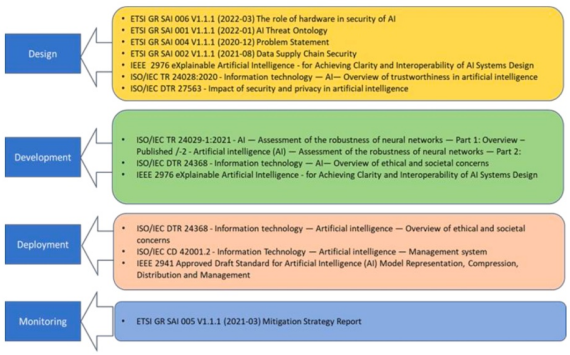

Figure 2: Specific standards used in the lifecycle of AI systems (ENISA, 2023).

To validate and evaluate the proposed methodology, a variety of standards will be utilized, including ISO/IEC WD 27090 (Guidance for addressing security threats to AI systems), ISO/IEC WD 27091 (AI - Privacy protection), ISO/IEC 27115 (Cybersecurity evaluation of complex systems), ISO/IEC

CD TR 27563 (Impact of security and privacy in AI use cases), ISO/IEC FDIS 42001 (AI management system), ISO/IEC 23894 (AI guidance on risk management), ISO/IEC 5259 series (Data quality for analytics and machine learning), ISO/IEC 24029 series (Assessment of the robustness of neural networks), ISO/IEC 22989 (AI concepts and terminology), and ISO/IEC 5338 (AI system lifecycle processes).

The study of these standards will enable the definition of a comprehensive knowledge framework for an AI system security optimization ecosystem, which integrates both risk assessment and human factors. By leveraging this ecosystem, organizations will be better equipped to understand and mitigate the complex socio-technical risks associated with adversarial ML attacks, thereby enhancing the overall security and trustworthiness of AI systems.

CONCLUSION

This paper highlights the urgent need for enhanced security frameworks that incorporate human factors into risk assessments for AI systems, particularly in response to the growing sophistication of adversarial ML attacks. It proposes an AI System Security Optimization Ecosystem that integrates insights from cybersecurity, cyberpsychology, and AI to address both technical and socio-technical aspects of security. By employing digital clones, XAI techniques, and innovative conversational agents, the ecosystem enhances the protection, transparency, and trustworthiness of AI systems. This approach not only improves security against various adversarial attacks but also advances the development of more robust, trustworthy AI technologies capable of operating securely in diverse environments.

ACKNOWLEDGMENT

The authors would like to acknowledge the financial support provided for the following projects: The ‘Collaborative, Multi-modal, and Agile Professional Cybersecurity Training Program for a Skilled Workforce in the European Digital Single Market and Industries’ (CyberSecPro) project, which has received funding from the European Union’s Digital Europe Programme (DEP) under grant agreement No. 101083594; the ‘Human-centered Trustworthiness Optimization in Hybrid Decision Support’ (THEMIS 5.0) project, which has received funding from the European Union’s Horizon Programme under grant agreement No. 101121042; the ‘Advanced Cybersecurity Awareness Ecosystem for SMEs’ (NERO) project, which has received funding from the European Union’s DEP programme under grant agreement No. 101127411; the ‘Harmonizing People, Processes, and Technology for Robust Cybersecurity’ (CyberSynchrony) project, which has received funding from the European Union’s Digital Europe Programme (DEP) under grant agreement No. 101158555; and the ‘Fostering Artificial Intelligence Trust for Humans towards the Optimization of Trustworthiness through Large-scale Pilots in Critical Domains’ (FAITH) project, which has received funding from the European Union’s Horizon Programme under grant agreement No. 101135932. RGL is also supported by the EU

Horizon2020 project MariCybERA under grant agreement No. 952360. The views expressed in this paper represent only the views of the authors and not of the European Commission or the partners in the above-mentioned projects. Finally, the authors declare that there are no conflicts of interest, including any financial or personal relationships, that could be perceived as potential conflicts.

REFERENCES

Biggio, B., & Roli, F. (2018). Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. Pattern Recognition, 84, 317–331.

Demetrio, L., Biggio, B., Lagorio, G., Zizzo, G., & Roli, F. (2021). Adversarial Machine Learning: A Systematic Review of Backdoor Attack and Defense. IEEE Transactions on Neural Networks and Learning Systems, 32(10), 4182–4205.

Doshi-Velez, F., & Kim, B. (2017). Towards a Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

Doshi-Velez, F., & Kim, B. (2020). Towards a Rigorous Science of Interpretable Machine Learning. Nature Machine Intelligence, 2(4), 222–230.

Dyrmishi, S., Ghamizi, S., & Cordy, M. (2023). How do humans perceive adversarial text? A reality check on the validity and naturalness of word-based adversarial attacks. ArXiv.

European Committee for Standardization (CEN) and European Committee for Electrotechnical Standardization (CENELEC). (2021). CEN/CLC/JTC 21 AI Risk Management framework. Joint Technical Committee document.

European Union Agency for Cybersecurity (ENISA). (2021). ENISA AI Cybersecurity Challenges: Threat Landscape for Artificial Intelligence. Available at: https://www.enisa.europa.eu/publications/artificial-intelligence-cybersecurity-challenges

Fogg, B. J. (2019). Tiny Habits: The Small Changes That Change Everything. Houghton Mifflin Harcourt.

Gilliard, Ezekia, Jinshuo Liu, and Ahmed Abubakar Aliyu. "Knowledge graph reasoning for cyber attack detection." IET Communications (2024).

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M. & Perez, E. (2024). Sleeper agents: Training deceptive llms that persist through safety training. arXiv preprint arXiv:2401.05566.

International Organization for Standardization (ISO). (2013). ISO/IEC 27001:2013 Information technology – Security techniques – Information security management systems – Requirements. Available at: https://www.iso.org/standard/54534.html

International Organization for Standardization (ISO). (2018). ISO 31000:2018 Risk management – Guidelines. Available at: https://www.iso.org/standard/65694.html

International Organization for Standardization (ISO). (2020). ISO/IEC WD 27090 Information security, cybersecurity and privacy protection – Framework for securing artificial intelligence systems. Draft document.

International Organization for Standardization (ISO). (2020). ISO/IEC WD 27091 Information security, cybersecurity and privacy protection – Guidelines for the management of AI system supply chain risks. Draft document.

International Organization for Standardization (ISO). (2022). ISO/IEC 22989:2022 Information technology – Artificial intelligence – Concepts and terminology. Available at: https://www.iso.org/standard/73822.html

International Organization for Standardization (ISO). (2022). ISO/IEC FDIS 5338 Information technology – Artificial intelligence – Guidelines for AI risk management. Final draft.

Li, S., Liu, H., Dong, T., Zhao, B. Z. H., Xue, M., Zhu, H., & Lu, J. (2021). Hidden Backdoors in Human-Centric Language Models. Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security.

Mao, X., Chen, Y., Wang, S., Su, H., He, Y., & Xue, H. (2020). Composite adversarial attacks. ArXiv.

National Institute of Standards and Technology (NIST). (2021). NIST Artificial Intelligence Risk Management Framework (AI RMF). Available at: https://www.nist.gov/artificial-intelligence/ai-risk-management-framework

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2016). SoK: Towards the Science of Security and Privacy in Machine Learning. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), 399–414.

Papernot, N., Mcdaniel, P., Goodfellow, I. J., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical black-box attacks against machine learning. Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security.

Pierazzi, F., Pendlebury, F., Cortellazzi, J., & Cavallaro, L. (2019). Intriguing properties of adversarial ML attacks in the problem space. 2020 IEEE Symposium on Security and Privacy (SP).

Sanders, E. B. N., & Stappers, P. J. (2019). Co-creation and the New Landscapes of Design. CoDesign, 4(1), 5–18.

Sarker, I. H., Janicke, H., Mohsin, A., Gill, A., & Maglaras, L. (2024). Explainable AI for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. ICT Express.

Sun, W., Jiang, X., Dou, S., Li, D., Miao, D., Deng, C., & Zhao, C. (2022). Invisible Backdoor Attack With Dynamic Triggers Against Person Re-Identification. ArXiv, abs/2211.10933.

Yang, W., Lin, Y., Li, P., Zhou, J., & Sun, X. (2021). Rethinking Stealthiness of Backdoor Attack against NLP Models. Proceedings of the 2021 Conference of the Association for Computational Linguistics.