International Journal of Internet, Broadcasting and Communication Vol.16 No.4 375-384 (2024)
https://doi.org/10.7236/IJIBC.2024.16.4.375

IJIBC 24-4-43

A Comprehensive Review of AI Security: Threats, Challenges, and Mitigation Strategies

Serdar Yazmyradov[1], Hoon Jae Lee[2*]

[1]*Department of Computer Engineering, Dongseo University, Busan, Korea*
[2]*Professor of Department Information Security of Dongseo University, Korea*
e-mail serikyazmuradov@gmail.com, hjlee@dongseo.ac.kr

Abstract

As Artificial Intelligence (AI) continues to permeate various sectors such as healthcare, finance, and transportation, the importance of securing AI systems against emerging threats has become increasingly critical. The proliferation of AI across these industries not only introduces opportunities for innovation but also exposes vulnerabilities that could be exploited by malicious actors. This comprehensive review delves into the current landscape of AI security, providing an in-depth analysis of the threats, challenges, and mitigation strategies associated with AI technologies. The paper discusses key threats such as adversarial attacks, data poisoning, and model inversion, all of which can severely compromise the integrity, confidentiality, and availability of AI systems. Additionally, the paper explores the challenges posed by the inherent complexity and opacity of AI models, particularly deep learning networks. The review also evaluates various mitigation strategies, including adversarial training, differential privacy, and federated learning, that have been developed to safeguard AI systems. By synthesizing recent advancements and identifying gaps in existing research, this paper aims to guide future efforts in enhancing the security of AI applications, ultimately ensuring their safe and ethical deployment in both critical and everyday environments.

Keywords: *AI security, AI healthcare, AI finance, AI Importance, Mitigation strategies*.

1. INTRODUCTION

Artificial Intelligence (AI) has transitioned from a niche academic pursuit to a transformative technology driving innovation across various industries, including healthcare, finance, and transportation. [33] The growing reliance on AI in these sectors underscores the critical need for robust security measures to protect AI systems from emerging threats. As AI systems become more integral to decision-making processes, the potential risks associated with their vulnerabilities also increase significantly.

The security of AI systems is a multifaceted challenge that involves not only technical considerations but also ethical, privacy, and trust issues. For instance, adversarial attacks, where subtle manipulations to input data can lead AI models to make incorrect predictions, highlight the potential dangers of compromised AI

Manuscript received: October. 21, 2024 / Revised: October. 26, 2024 / Accepted: October. 31, 2024
Corresponding Author: hjlee@dongseo.ac.kr
Tel: ***-****-**** Fax: 051-313-1046
Author's affiliation (Professor, Department of Information Security, Dongseo University, Korea)

systems. [1,3,34] Such attacks can be particularly devastating in high-stakes environments, such as autonomous driving or medical diagnostics, where erroneous outputs can have severe consequences. Similarly, data poisoning, which involves the injection of malicious data into the training phase of AI models, poses significant threats to the integrity and reliability of AI-driven decisions. Additionally, model inversion and extraction attacks can lead to the exposure of sensitive information and the theft of proprietary models, raising serious concerns about privacy and intellectual property. [38,39] Compounding these issues is the inherent opacity of many AI models, especially deep learning networks, which often function as "black boxes" with decision-making processes that are difficult to interpret. [5,6] This lack of transparency complicates efforts to detect and mitigate security breaches, as understanding the internal workings of AI systems is crucial for identifying and addressing vulnerabilities. Furthermore, the dynamic and rapidly evolving nature of AI technologies necessitates continuous adaptation of security measures to counter new and emerging threats. In response to these challenges, researchers have developed a range of mitigation strategies aimed at enhancing AI security. Techniques such as adversarial training, which strengthens models by exposing them to adversarial examples during training, and differential privacy, which protects individual data points from being exposed, are among the strategies designed to fortify AI systems. [2] Federated learning, which allows models to be trained across decentralized devices while keeping data localized, offers another promising approach to preserving privacy and security in AI applications. [36] However, despite these advancements, there remain significant gaps in our understanding and implementation of effective AI security measures. [37]

This review paper seeks to provide a comprehensive analysis of the current state of AI security by examining the major threats, challenges, and mitigation strategies in this field. Through a synthesis of recent research, this paper aims to identify critical areas for further investigation and contribute to the development of more secure, robust, and trustworthy AI systems. The ultimate objective is to ensure that as AI continues to advance and integrate into critical sectors, it does so in a manner that is both safe and ethical, safeguarding the interests of individuals, organizations, and society.

Background Artificial Intelligence (AI) encompasses a wide range of technologies such as machine learning (ML), deep learning (DL), and natural language processing (NLP) that enable systems to learn from data and make decisions. The growing reliance on AI across different sectors necessitates robust security measures to safeguard these systems against various threats.

Importance of AI Security AI security is critical due to the widespread use of AI in sensitive and mission-critical applications. Security breaches can lead to significant consequences, including financial losses, privacy violations, and even threats to human safety.

Objectives of the Review This review aims to provide a detailed examination of the current state of AI security, highlighting the various threats and challenges, and discussing the strategies and solutions developed to mitigate these risks.

2. THREATS TO AI SYSTEMS

AI systems, while powerful, are not immune to various security threats that can compromise their functionality and reliability. Among these, adversarial attacks are particularly insidious, as they involve subtle manipulations of input data that can deceive AI models into making incorrect predictions. These attacks are of significant concern, especially in critical applications like autonomous vehicles and healthcare, where erroneous outputs can have severe consequences.

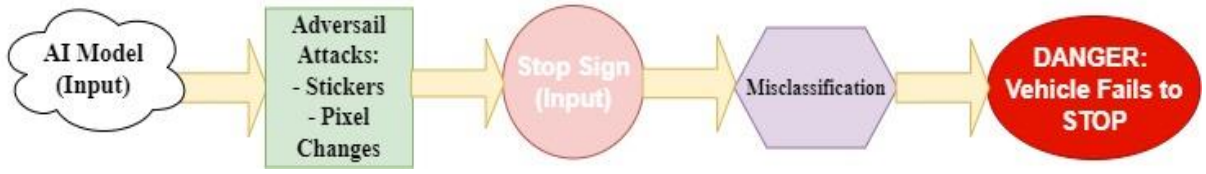

Figure 1. This figure outlines the process of how adversarial attacks can impact AI models, using a scenario involving an autonomous vehicle and a stop sign to illustrate the concept.

2.1 Adversarial Attacks

Adversarial attacks exploit vulnerabilities in AI models, particularly in machine learning systems, by introducing small, often imperceptible changes to input data that cause the model to make incorrect or unexpected decisions. These attacks challenge the robustness of AI systems and underscore the need for more resilient models. For example, in an AI-powered image recognition system designed to identify stop signs in an autonomous vehicle, the system would typically recognize a stop sign and prompt the vehicle to halt. However, during an adversarial attack, an attacker could make subtle modifications—such as placing strategically positioned stickers or altering pixel values on the stop sign—causing the AI model to misclassify it as something else, like a yield sign or a speed limit sign. This misclassification could prevent the vehicle from stopping, potentially leading to catastrophic consequences.

2.1.1 Types of Adversarial Attacks

Fast Gradient Sign Method (FGSM): FGSM is one of the simplest and most well-known adversarial attack methods. It perturbs the input data by adding a small amount of noise in the direction of the gradient of the loss function with respect to the input. This perturbation is calculated to maximize the model's prediction error while remaining imperceptible to human observers. For example, consider an image of a cat that is correctly classified by an AI model. FGSM might add a small amount of noise to the image, altering pixel values just enough so that the model now misclassifies the cat as a dog. This change would be imperceptible to the human eye but sufficient to fool the AI model. [1]

Projected Gradient Descent (PGD): Projected Gradient Descent (PGD) is an iterative extension of the Fast Gradient Sign Method (FGSM) and is considered one of the most potent first-order adversarial attacks. In this approach, small perturbations are applied repeatedly to the input data, with each iteration fine-tuning the attack to enhance its effectiveness. After each adjustment, the perturbation is projected back into a predefined constraint set to ensure it stays within a specific range. For instance, starting with an image of the handwritten digit '7', PGD could iteratively modify the pixels, gradually leading the model to classify the digit as '1'. Although these modifications are subtle and kept within a minimal range to avoid detection, the resulting adversarial example may still look like a '7' to a human observer, while the model incorrectly recognizes it as a '1'.[2]

Carlini & Wagner (C&W) Attacks: Carlini & Wagner (C&W) attacks are among the most effective and sophisticated adversarial techniques. They work by optimizing a loss function that minimizes the perturbation while ensuring that the AI model's prediction is incorrect. These attacks are particularly dangerous because they can create adversarial examples that are visually indistinguishable from the original inputs but are highly effective at deceiving the model. For example, in a facial recognition system, a C&W attack might introduce subtle changes to an image of Person A, causing the AI model to incorrectly identify the image as Person B. Despite these alterations, the image would still appear as Person A to the human eye, making it possible for such an attack to bypass security systems that rely on facial recognition.[3]

2.1.2 Impact of Adversarial Attacks

Adversarial attacks can have severe impacts, particularly in areas where AI systems are deployed in safety-critical applications. The ability of an attacker to deceive an AI model can lead to catastrophic outcomes, as illustrated by the following examples:

- **Autonomous Vehicles**: Adversarial attacks on image recognition systems in autonomous vehicles can cause the misclassification of road signs, leading to incorrect decisions like failing to stop at a stop sign. Demonstrated that placing stickers on stop signs could cause an AI model to misclassify them as speed limit signs, posing significant safety risks. [4]
- **Healthcare Systems**: In medical imaging, adversarial attacks could alter scans or images just enough to cause a diagnostic AI model to misinterpret a benign tumor as malignant or vice versa. This could lead to incorrect treatment decisions, potentially endangering patients' lives. [13]
- **Security Systems**: In facial recognition systems used for security, adversarial attacks could allow unauthorized individuals to gain access to secure areas by manipulating images so that the AI model misidentifies them as authorized personnel. [14]

3. CHALLENGES IN AI SECURITY

As AI systems become increasingly prevalent in various sectors, addressing the challenges associated with their security and reliability has become paramount. This section explores three critical challenges: explainability and transparency, robustness and reliability, and ethical and legal issues. Each of these areas presents significant hurdles that must be overcome to ensure the safe and responsible deployment of AI technologies.

3.1 Explainability and Transparency

One of the most pressing challenges in AI is the lack of explainability and transparency, particularly in complex models such as deep learning networks. These models often function as "black boxes," where the decision-making processes are opaque and difficult to interpret. This lack of transparency can hinder trust and accountability in AI systems, especially in high-stakes environments like healthcare, finance, and criminal justice. [5]

The black-box nature of many AI models limits their interpretability [9], making it challenging for users to understand the reasoning behind specific predictions or decisions. This issue is especially critical in domains where transparency is essential for accountability and trust. Also emphasize the importance of interpretability in AI [6], proposing that transparent models are crucial for ensuring that AI systems are used responsibly and ethically. Furthermore, provide a comprehensive survey on methods for enhancing the interpretability of AI models, discussing approaches such as model simplification, visualization techniques, and post-hoc explanations. [7] For example, in healthcare, an AI model might be used to predict the likelihood of a patient developing a certain disease. However, if the model's decision-making process is not transparent, it can be challenging for doctors to trust the prediction, especially when it contradicts clinical intuition. This lack of explainability can lead to skepticism and underutilization of AI technologies in critical areas.

3.2 Robustness and Reliability

Ensuring that AI systems are robust and reliable under various conditions is a significant challenge. AI models can be highly sensitive to small changes in input data, leading to incorrect or unexpected outputs. This

vulnerability is especially concerning in safety-critical applications, where errors can have severe consequences.

Highlight the susceptibility of AI models to adversarial attacks and emphasize the importance of developing robust models that can resist such manipulations. [8] Similarly, discuss the challenges of creating AI systems that perform reliably across diverse and potentially adversarial environments, suggesting that improving robustness requires not only technical solutions but also a deeper understanding of the principles of machine learning. [1] Further examine the limitations of current defenses against adversarial attacks, underscoring the need for more effective strategies to ensure the reliability of AI systems. [3] For instance, in autonomous vehicles, AI systems must accurately recognize and respond to various road signs and conditions. However, even minor perturbations, such as graffiti on a stop sign, can cause the AI to misinterpret the sign, potentially leading to dangerous situations. Ensuring robustness in such systems is critical to their safe deployment. [12]

3.3 Ethical and Legal Issues

The deployment of AI systems presents numerous ethical and legal challenges, including concerns about privacy, bias, accountability, and the potential for misuse. To ensure fair and responsible use, AI systems must be designed and implemented in ways that adhere to ethical standards and comply with legal regulations.

The ethical implications of AI decision-making, particularly regarding fairness and bias. He argues that without proper safeguards, AI systems can perpetuate existing inequalities and reinforce discriminatory practices. Similarly, explore the global landscape of AI ethics, highlighting the need for consistent ethical frameworks to guide the development and deployment of AI technologies. Emphasize the dual-use nature of AI, stressing the importance of establishing legal and regulatory frameworks to prevent the misuse of AI technologies in harmful or malicious ways. For example, AI algorithms used in hiring processes have come under scrutiny for potentially perpetuating bias and discrimination. If an AI system is trained on biased data, it may favor certain demographic groups over others, leading to unfair hiring practices. Ensuring that AI systems are both ethical and legally compliant is essential to avoid such issues. [9,10,11]

4. MITIGATION STRATEGIES

In securing data, effective mitigation strategies are crucial for protecting against unauthorized access, data breaches, and other security risks. These strategies focus on maintaining data integrity, confidentiality, and availability, while addressing potential challenges.

Table 1. ***This table summarizes key mitigation strategies for securing data, along with objectives, strategies, and a comprehensive overview of each aspect.***

Category	Objective	Strategies
1. Data Integrity	Ensure data accuracy and reliability throughout its lifecycle.	- Implement regular integrity checks using cryptographic hash functions and checksums. - Use data validation techniques to detect and correct errors. [22]
2. Data Confidentiality	Protect data from unauthorized access to ensure privacy and prevent breaches.	- Utilize strong encryption methods for data at rest and in transit. - Apply data anonymization techniques such as masking and tokenization. [24,25]

3. Data Availability	Ensure data accessibility for authorized users when needed without disruptions.	- Deploy redundant storage solutions and robust disaster recovery plans. - Use high-availability configurations and regular backups. [26,27]
4. Access Controls	Manage and restrict data access based on user roles and permissions.	- Implement multi-factor authentication (MFA) and role-based access control (RBAC). - Regularly review and update access permissions. [28,29]
5. Secure Storage	Protect data stored on cloud and on-premises systems from unauthorized access.	- Use encryption for stored data. - Adopt secure cloud services and ensure physical and logical security measures for on-premises storage. [28,30]
6. Integrity Checks	Verify that data has not been altered or tampered with.	- Use checksums, hash functions, and continuous monitoring. - Implement anomaly detection systems. [23, 31]

Table 2. ***This table summarizes the Challenges and Description perspectives.***

Challenges	Description
Complexity	Integrating and managing diverse security technologies can be complex. [28]
Resource Intensity	Significant investment in tools, technology, and expertise is required. [28]
Balancing Security and Usability	Finding a balance between stringent security measures and user convenience. [32]
Evolving Threats	Continuous adaptation to new and emerging threats is necessary. [28]
Compliance	Ensuring adherence to regulatory and industry standards adds complexity. [32]

5. FUTURE DIRECTIONS

As artificial intelligence (AI) continues to evolve, addressing its security challenges is increasingly critical. This section outlines key future directions to enhance AI security and resilience against adversarial threats, supported by relevant literature.

5.1 Advanced Defense Mechanisms

Developing advanced defense mechanisms is essential to counter the evolving nature of adversarial attacks on AI systems. As adversaries devise more sophisticated methods to exploit vulnerabilities, defense strategies must evolve correspondingly. Innovations in anomaly detection, adaptive security protocols, and resilient algorithm design are necessary to anticipate and mitigate potential threats effectively. Demonstrated the potential of adversarial examples in machine learning, highlighting the need for advanced defensive techniques [1]. Additionally, robust optimization methods to enhance model resilience against adversarial attacks [2].

5.2 Interdisciplinary Research

Collaborative efforts between AI researchers, cybersecurity experts, and legal scholars are vital in crafting comprehensive AI security frameworks. Such interdisciplinary research can bridge gaps between technical, legal, and ethical perspectives, leading to more robust and holistic solutions. For instance, emphasizes the importance of cross-disciplinary collaboration to address the multifaceted challenges of AI security [18]. By integrating diverse expertise, it is possible to address security challenges from multiple angles, ensuring that AI technologies are both effective and secure.

5.3 Policy and Regulation

Establishing clear policies and regulations is crucial for guiding the ethical deployment and use of AI technologies. Regulatory frameworks should address issues related to data privacy, algorithmic transparency, and accountability. Effective policies can provide guidelines for the responsible development and implementation of AI systems, ensuring that they are used in ways that are both secure and aligned with societal values. The necessity of regulatory frameworks for ethical AI deployment, advocating for transparent and accountable practices [9]. Furthermore, the European Union's General Data Protection Regulation (GDPR) represents a significant step towards ensuring data protection and privacy in AI applications [19].

5.4 Explainable AI (XAI)

Enhancing the explainability of AI systems, also known as Explainable AI (XAI), is a key factor in improving transparency and trustworthiness. By making AI decision-making processes more understandable to users and stakeholders, XAI can help identify and address potential security concerns. Explainability not only fosters trust but also facilitates the detection of vulnerabilities and biases within AI models. The LIME technique, which provides local interpretability of AI models, contributing to improved transparency [20]. Additionally, outline various methods and challenges in developing explainable AI systems [6].

5.5 Continuous Monitoring

Implementing continuous monitoring systems for AI applications is critical for real-time detection and response to security incidents. These systems can provide ongoing surveillance of AI operations, enabling the prompt identification of anomalies or suspicious activities. Real-time monitoring helps in rapidly addressing security threats, minimizing potential damage, and maintaining the integrity of AI systems. The highlight the importance of continuous monitoring and anomaly detection in maintaining AI security, emphasizing the need for adaptive and responsive systems [21].

In summary, advancing AI security requires a multi-faceted approach that includes developing sophisticated defense mechanisms, fostering interdisciplinary collaboration, establishing clear policies, enhancing explainability, and implementing continuous monitoring. By pursuing these directions, we can better safeguard AI technologies against emerging threats and ensure their secure and ethical deployment.

6. DISCUSSION

This study differentiates itself from other reviews in the field of AI security in several key aspects. Firstly, while comprehensively addressing the existing literature on AI security, this study takes a holistic perspective on the threats, challenges, and prevention strategies within the context of AI security. The study focuses not

only on evaluating threats to AI security through attacks and technical vulnerabilities but also on a broader range of issues such as complexity, lack of transparency, and ethical problems associated with AI systems. By highlighting the AI security problems in critical sectors such as healthcare, finance, and autonomous vehicles, it addresses the fact that potential threats are not only technical but also ethical and social in nature.

In this context, one of the most notable distinctions of this study from others is its multi-faceted analysis of the topic of AI security. The study categorizes threats to AI systems into adversarial attacks, data poisoning, and model reverse engineering, while also considering the sectoral differences. Additionally, it relates security issues to the incomprehensible nature of complex AI models, emphasizing that the lack of transparency makes it difficult to detect and respond to security breaches. This situation fills a significant gap that has not been adequately addressed in other reviews and provides a roadmap to enhance the reliability and security of AI systems.

Furthermore, this study delves deeply into the proposed preventive strategies for AI security and questions the effectiveness of defense techniques, aiming to offer advanced solutions. The effects of techniques such as adversarial training, differential privacy, and federated learning are detailed, and the applicability and weaknesses of these strategies are examined. By addressing the shortcomings of these preventive strategies, the study emphasizes the need for their improvement and presents a framework for future research in the field of AI security.

Finally, this review positions itself differently in the literature by addressing future trends related to AI security. The study presents recommendations such as advanced defense mechanisms, continuous monitoring systems, and the development of policies and regulations, stating that AI security must be continually updated to keep pace with the rapidly evolving nature of this field. Notably, emphasizing the importance of ensuring transparency in AI systems through XAI (Explainable AI) applications for security and reliability underscores the distinctiveness of this study.

7. CONCLUSION

Artificial intelligence (AI) security is a vital research area in the rapidly advancing field of technology. While AI offers immense benefits and opportunities, ensuring the security of these systems is equally essential. AI systems are prone to errors with unpredictable outcomes, vulnerable to malicious attacks, and capable of making decisions that might lead to ethical concerns. These risks highlight the importance of continuous research and the development of innovative solutions to address emerging threats in AI security. Understanding the threats and challenges associated with AI is fundamental to developing effective security strategies. Key threats include data manipulation, attacks on AI models, and errors that may cause systems to behave unpredictably. By conducting thorough threat analyses and implementing strong risk assessments, we can enhance the security of AI systems and prevent their misuse. Additionally, the ethical use and transparency of AI are crucial components of any security approach. Ensuring that AI decision-making processes are transparent and understandable is essential for the secure and responsible use of these technologies.In conclusion, AI security is not just a technical issue but also one with significant social and ethical implications. The safe and responsible use of AI technologies will require strong security measures and ongoing innovation. Researchers, policymakers, and industry leaders must work together to mitigate potential risks while maximizing the benefits of AI. Through this collaborative effort, we can ensure that AI reaches its full potential and delivers positive outcomes for society as a whole.

ACKNOWLEDGEMENT

This thesis was supported by 'The Construction Project for Regional Base Information Security Cluster', grant funded by Ministry of Science, ICT and Busan Metropolitan City in 2024.

REFERENCES

[1] Goodfellow, I., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. https://doi.org/10.48550/arXiv.1412.6572

[2] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards Deep Learning Models Resistant to Adversarial Attacks. https://doi.org/10.48550/arXiv.1706.06083

[3] N. Carlini and D. Wagner, "Towards Evaluating the Robustness of Neural Networks," *2017 IEEE Symposium on Security and Privacy (SP)*, San Jose, CA, USA, 2017, pp. 39-57, doi: 10.1109/SP.2017.49. keywords: {Neural networks;Robustness;Measurement;Speech recognition;Security;Malware;Resists}

[4] Eykholt, K., Evtimov, I., Fernandes, E., et al. (2018). Robust Physical-World Attacks on Deep Learning Models.

[5] Rudin, C. (2019). Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9122117/

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

[7] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42. https://doi.org/10.1145/3236009

[8] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*. https://doi.org/10.48550/arXiv.1706.06083

[9] Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. *Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency*.

[10] Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399.

[11] Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Anderson, H. (2018). The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*. https://doi.org/10.48550/arXiv.1802.07228

[12] Farnoosh Heidarivincheh, Majid Mirmehdi, Dima Damen "Weakly-Supervised Completion Moment Detection using Temporal Attention" https://doi.org/10.48550/arXiv.1910.09920

[13] Finlayson, S. G., Bowers, J. D., Ito, J., Zittrain, J. L., Beam, A. L., & Kohane, I. S. (2019). "Adversarial attacks on medical machine learning." *Science*, 363(6433), 1287-1289. DOI: 10.1126/science.aaw4399

[14] Sharif, M., Bhagavatula, S., Bauer, L., & Reiter, M. K. (2016). "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. https://doi.org/10.1145/2976749.2978392

[15] Moosavi-Dezfooli, S.M., et al. (2016). "DeepFool: A simple and accurate method to fool deep neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Described an iterative method for generating minimal adversarial perturbations.

[16] Moosavi-Dezfooli, S.M., et al. (2017). "Universal adversarial perturbations." *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (CVPR). Proposed universal adversarial perturbations effective across different inputs.

[17] Papernot, N., et al. (2016). "The limitations of deep learning in adversarial settings." *IEEE European Symposium on Security and Privacy (EuroS&P)*. Introduced the JSMA method and discussed its effectiveness and limitations.

[18] Solon, O. (2020). How AI's Interdisciplinary Research Approach Can Improve Security. *The Guardian*.

[19] European Union. (2018). General Data Protection Regulation (GDPR). *Official Journal of the European Union*.

[20] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?" Explaining the Predictions of Any Classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. https://doi.org/10.1145/2939672.2939778

[21] Zhang, X., Zheng, Y., & Yang, H. (2020). Continuous Monitoring for Anomaly Detection in AI Systems. *IEEE Transactions on Knowledge and Data Engineering*.

[22] S. Vaudenay, "On the Security of Encryption Schemes," in *Advances in Cryptology - CRYPTO 2001*, Springer, 2001.

[23] Menezes, A.J., van Oorschot, P.C., & Vanstone, S.A. (1997). Handbook of Applied Cryptography (1st ed.). CRC Press. https://doi.org/10.1201/9780429466335

[24] W. Stallings, *Computer Security: Principles and Practice*, Pearson, 2020.

[25] D. Kahn, *The Codebreakers: The Comprehensive History of Secret Communication from Ancient Times to the Internet*, Scribner, 1996.

[26] M. G. Schwartz, *Principles of Computer System Design: An Introduction*, Morgan Kaufmann, 2004.

[27] G. Coulouris, J. Dollimore, and T. Kindberg, *Distributed Systems: Concepts and Design*, Addison-Wesley, 2011.

[28] R. Anderson, *Security Engineering: A Guide to Building Dependable Distributed Systems*, Wiley, 2020.

[29] J. C. Brustoloni, "Role-Based Access Control in Large-Scale Distributed Systems," in *Proceedings of the IEEE Symposium on Security and Privacy*, 2005.

[30] P. Mell and T. Grance, *The NIST Definition of Cloud Computing*, National Institute of Standards and Technology, 2011.

[31] S. M. Bellovin and W. R. Cheswick, *Firewalls and Internet Security: Repelling the Wily Hacker*, Addison-Wesley, 2003.

[32] S. H. Kim and D. S. Kim, "Challenges and Strategies in Data Security and Privacy," *International Journal of Information Security*, vol. 12, no. 2, pp. 103-115, 2013.

[33] Goodfellow, I., Bengio, Y., & Courville, A. (2016). "Deep Learning." MIT Press.

[34] Szegedy, C., et al. (2014). "Intriguing Properties of Neural Networks." arXiv preprint.

[35] Zachary C. Lipton “The Mythos of Model Interpretability” https://doi.org/10.48550/arXiv.1606.03490

[36] McMahan, B., et al. (2017). "Communication-Efficient Learning of Deep Networks from Decentralized Data." AISTATS.

[37] Wang, L., et al. (2020). "Supply Chain Risks in AI: Identification and Mitigation." ACM Computing Surveys.

[38] Fredrikson, M., et al. (2015). "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures." ACM CCS.

[39] Shokri, R., et al. (2017). "Membership Inference Attacks Against Machine Learning Models." IEEE S&P.