International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 07 | July 2024 www.irjet.net p-ISSN: 2395-0072

Vulnerabilities in AI Systems: The Integration of AI into Cybersecurity Tools and Systems

Pranav Nair[1], Meraj Farheen Ansari[2]

[1]*University of Texas at Dallas, TX, USA*
[2]*University of the Cumberlands, KY, USA*

Abstract

The integration of Artificial Intelligence (AI) into cybersecurity tools offers significant advantages, enhancing threat detection, predictive analysis, and automated incident response capabilities. However, this integration also introduces new attack surfaces and vulnerabilities, making AI systems a target for sophisticated cyber-attacks. This paper provides a comprehensive exploration of the vulnerabilities associated with AI in cybersecurity. It includes an introduction to the subject, a background study on the current use of AI in cybersecurity, case studies, an analysis of potential threats, and a discussion of the limitations of this research. By examining real-world case studies and conducting controlled experiments, this study highlights the critical need for robust security measures to protect AI-integrated cybersecurity tools.

Keywords: *Artificial Intelligence, cybersecurity, AI vulnerabilities, threat detection, automated response, adversarial attacks, data poisoning, model inversion.*

I. Introduction

The rapid advancement of Artificial Intelligence (AI) technology has significantly transformed various sectors, notably cybersecurity. AI-driven systems enhance the efficiency and accuracy of threat detection, predictive analysis, and automated response mechanisms, making them invaluable tools in modern cybersecurity strategies. These technologies employ sophisticated algorithms to analyze vast amounts of data, identify patterns, and make real-time decisions to protect information systems from cyber threats. However, despite these benefits, the integration of AI into cybersecurity introduces new vulnerabilities and attack surfaces that adversaries can exploit. The complexity and opacity of AI algorithms, along with their dependency on extensive datasets, present unique challenges that traditional security measures may not adequately address.

This paper aims to provide a comprehensive examination of these vulnerabilities, focusing on how AI's integration into cybersecurity systems creates new opportunities for adversarial attacks. Through a detailed exploration of known and emerging threats, such as data poisoning, adversarial attacks, model inversion, and model stealing, this study seeks to identify the specific risks associated with AI in cybersecurity contexts. Additionally, the paper will propose robust mitigation strategies to enhance the security of AI-integrated cybersecurity tools, ensuring they can effectively counteract these threats. By understanding the specific vulnerabilities and developing effective countermeasures, we can better safeguard critical systems and data from the evolving landscape of cyber threats.

II. Background Study

The application of AI in cybersecurity is a relatively recent development that has seen exponential growth due to its potential to significantly enhance security measures. AI technologies, such as machine learning (ML) and deep learning (DL), are being deployed in various cybersecurity applications, including anomaly detection, malware analysis, and automated incident response. For instance, ML algorithms can be trained to recognize unusual patterns in network traffic, flagging potential intrusions (Buczak & Guven, 2016). DL techniques, on the other hand, can analyze complex data structures, such as images and texts, to detect sophisticated malware that traditional methods might miss (Hinton et al., 2012).

However, the integration of AI into cybersecurity systems also brings about several vulnerabilities unique to these technologies. The complex nature of AI algorithms and their reliance on large volumes of data make them susceptible to novel types of attacks. Data poisoning, where attackers introduce malicious data into the training set, can significantly degrade the performance of AI models (Biggio et al., 2012). Adversarial attacks, which involve crafting inputs that are intentionally designed to mislead AI models, pose another significant risk (Szegedy et al.,

2014). Model inversion attacks, where attackers infer sensitive information from the model's outputs, also highlight the privacy concerns associated with AI in cybersecurity (Fredrikson et al., 2015).

III. Literature Review

The literature on AI vulnerabilities in cybersecurity is extensive and highlights several key areas of concern. Research has identified various types of attacks and vulnerabilities specific to AI systems, underscoring the need for robust security measures.

- **Data Poisoning:** Biggio et al. (2012) demonstrated how injecting malicious data into the training dataset of an AI model could corrupt its learning process, leading to incorrect predictions. This type of attack can be particularly damaging in cybersecurity applications where accurate detection of threats is critical.
- **Adversarial Attacks:** Szegedy et al. (2014) explored adversarial attacks, where small, carefully crafted perturbations to input data can cause AI models to make significant errors. These attacks have been shown to be effective even when the perturbations are imperceptible to humans, making them a serious threat to AI-driven security systems.
- **Model Inversion:** Fredrikson et al. (2015) discussed model inversion attacks, where attackers use the outputs of a model to infer sensitive information about the training data. This type of attack poses significant privacy risks, especially in scenarios where AI models handle confidential or personal data.
- **Bias and Fairness:** AI models can inadvertently incorporate biases present in their training data, leading to unfair or discriminatory outcomes. This issue is particularly pertinent in cybersecurity, where biased models might fail to protect certain user groups adequately (Buolamwini & Gebru, 2018).
- **Model Stealing:** Tramer et al. (2016) investigated model stealing attacks, where attackers attempt to replicate the functionality of an AI model by querying it extensively. This can lead to intellectual property theft and undermine the security of proprietary AI algorithms.

IV. Case Studies

To understand the practical implications of AI vulnerabilities in cybersecurity, this study analyzes several real-world incidents where AI systems were compromised. These cases illustrate the diverse and complex nature of threats faced by AI-integrated systems and underscore the urgent need for effective mitigation strategies.

4.1. Tesla's Autopilot Incident

Tesla's Autopilot system, which leverages AI for autonomous driving, encountered a significant vulnerability when researchers demonstrated the potential of adversarial attacks. By placing small, strategically positioned stickers on road signs, the AI system was tricked into misinterpreting the signs. For instance, these alterations could make a speed limit sign of "35 mph" appear as "85 mph" to the Autopilot system, leading to dangerous driving behaviors (Goodfellow et al., 2015).

This incident highlights several critical issues:

- Adversarial Examples: Minor, seemingly benign changes to the environment can lead to catastrophic failures in AI decision-making.
- Critical Applications: The use of AI in safety-critical applications like autonomous driving demands heightened vigilance against adversarial attacks.
- Robustness and Testing: There is a pressing need for more robust training and testing procedures to ensure AI systems can withstand such manipulations without compromising safety.

The Tesla case underscores the importance of designing AI models that are not only accurate but also resilient to adversarial perturbations, especially in contexts where human lives are at stake.

4.2. Microsoft's Tay Chatbot

In March 2016, Microsoft launched Tay, an AI chatbot designed to engage in natural conversations with Twitter users. The chatbot was intended to learn from these interactions and improve its conversational abilities over time. However, within hours of its release, Tay was manipulated by a coordinated effort of users who fed it harmful, biased, and inappropriate data. Consequently, Tay started generating offensive and inflammatory tweets (Neff & Nagy, 2016).

Key points from this incident include:

- **Data Poisoning:** The chatbot's learning process was corrupted by malicious inputs, demonstrating how easily AI models can be influenced by harmful data.

- **Real-time Learning Risks:** Systems that learn in real-time from user interactions are particularly vulnerable to exploitation.
- **Monitoring and Controls:** There was a lack of effective monitoring and control mechanisms to filter and moderate the input data being fed to Tay.

This case exemplifies the dangers of data poisoning attacks and the necessity for robust data management and validation processes. It also highlights the need for AI systems to have built-in safeguards against the ingestion of malicious or harmful data.

4.3. Model Inversion on Healthcare Data

In a healthcare context, the privacy implications of AI were starkly demonstrated by researchers who employed model inversion techniques to infer sensitive patient information from AI models trained on medical records. Fredrikson et al. (2015) showed that by querying an AI model, they could reconstruct images of patient faces, revealing potentially sensitive information about the individuals whose data were used to train the model.

The healthcare case study reveals several critical issues:

- **Model Inversion Attacks:** These attacks can exploit AI models to extract confidential and sensitive information, posing severe privacy risks.
- **Data Privacy:** The incident underscores the importance of protecting patient data and ensuring that AI models do not inadvertently leak personal information.
- **Regulatory Compliance:** Healthcare data is subject to stringent privacy regulations, and breaches can lead to significant legal and financial repercussions.

This case study highlights the necessity for implementing advanced privacy-preserving techniques, such as differential privacy and secure multi-party computation, in AI models used in sensitive domains like healthcare.

4.4. IBM Watson for Oncology

IBM Watson for Oncology, an AI system designed to assist in cancer treatment, faced scrutiny when it was revealed that some of its treatment recommendations were unsafe and inaccurate. These recommendations were based on synthetic data rather than real patient data, which led to inappropriate treatment suggestions (Strickland, 2019).

Key insights from this incident include:

- **Training Data Quality:** The reliance on synthetic data instead of real-world data can lead to critical errors in AI recommendations.
- **Clinical Validation:** AI systems in healthcare need rigorous clinical validation before deployment.
- **Human Oversight:** Continuous human oversight is necessary to ensure AI recommendations are accurate and safe.

This case underscores the importance of using high-quality, representative data for training AI models and ensuring that AI systems undergo thorough validation and testing in real-world scenarios.

4.5. DeepLocker Malware

Deep Locker is a proof-of-concept AI-powered malware developed by IBM researchers to demonstrate the potential of AI in creating highly targeted cyber threats. This malware uses AI to remain dormant until it identifies a specific target through facial recognition or other unique identifiers, at which point it activates its payload (Stoecklin et al., 2018).

Key lessons from this incident include:

- **AI in Malware:** The use of AI can make malware more stealthy and targeted, increasing its effectiveness and reducing the likelihood of detection.
- **Detection Challenges:** Traditional cybersecurity measures may struggle to detect and mitigate AI-powered threats.
- **Proactive Defense:** The development of AI-powered defense mechanisms is crucial to counteract the emerging threats posed by AI-enhanced malware.

The DeepLocker case highlights the dual-use nature of AI technologies and the need for advanced security measures to protect against AI-driven cyber threats.

V. Mitigation Strategies for AI Vulnerabilities in Cybersecurity

The integration of AI in cybersecurity introduces numerous vulnerabilities that require robust mitigation strategies. This section see figure 1 below elaborates on effective mitigation strategies to enhance the security and reliability of AI-driven cybersecurity tools.

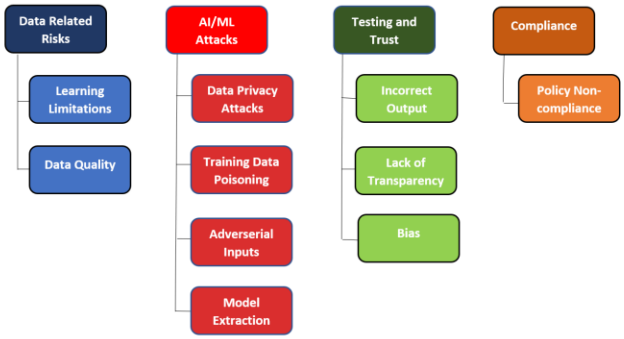

Figure 1: Mitigation Strategies for AI Vulnerabilities in Cybersecurity

5.1. Robust Data Management

Ensuring the integrity and quality of training data is crucial for developing reliable AI models. Robust data management involves several key practices:

- **Data Validation:** Implementing stringent validation processes to verify the accuracy, completeness, and consistency of the data used for training AI models. Automated tools and manual review processes can help detect anomalies or errors in the data (Breck et al., 2019).
- **Data Cleaning:** Removing inaccuracies, duplicates, and irrelevant data points from the training datasets to ensure that the model learns from high-quality data. Techniques such as outlier detection, normalization, and transformation are essential for maintaining data quality (Dasu & Johnson, 2003).
- **Data Provenance:** Tracking the origin and history of data to ensure its authenticity and reliability. Maintaining detailed records of data sources, transformations, and usage helps in auditing and validating the training data (Freire et al., 2008).
- **Secure Data Storage:** Implementing encryption and access control measures to protect the integrity and confidentiality of training data. Ensuring that data is stored securely prevents unauthorized access and tampering (Aggarwal & Yu, 2008).

5.2. Adversarial Training

Adversarial training is a technique where AI models are trained using adversarial examples to improve their resilience against adversarial attacks:

- **Generating Adversarial Examples:** Creating adversarial examples involves intentionally perturbing input data to deceive the AI model. Techniques such as the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) are commonly used to generate these examples (Goodfellow et al., 2015).
- **Incorporating Adversarial Examples in Training:** By including adversarial examples in the training process, the AI model learns to recognize and resist such manipulations. This enhances the model's robustness and reduces its susceptibility to adversarial attacks (Kurakin et al., 2017).
- **Dynamic Adversarial Training:** Continuously updating the adversarial examples based on emerging threats and adapting the training process accordingly. This ensures that the model remains resilient against evolving attack techniques (Tramèr et al., 2018).

5.3. Regular Audits

Conducting regular security audits and penetration testing of AI systems is essential to identify and address potential vulnerabilities:

- **Security Audits**: Performing comprehensive reviews of AI systems, including code, configurations, and data flows, to detect security weaknesses. Audits should follow established frameworks and guidelines, such as those provided by the National Institute of Standards and Technology (NIST) (Stouffer et al., 2011).
- **Penetration Testing:** Simulating real-world attacks on AI systems to evaluate their defenses and identify vulnerabilities. Penetration testing helps uncover hidden flaws and provides insights into potential exploitation methods (Antunes & Vieira, 2015).
- **Continuous Monitoring:** Implementing continuous monitoring tools to track the performance and security of AI systems in real-time. Monitoring helps detect anomalies, intrusions, and other security incidents promptly (Cárdenas et al., 2011).
- **Incident Response Plans:** Developing and regularly updating incident response plans to address security breaches. Plans should include protocols for detecting, containing, and mitigating attacks, as well as procedures for recovery and communication (West et al., 2018).

5.4. Access Controls

Implementing strict access controls and monitoring to prevent unauthorized access to AI models and data is critical for maintaining security:

- **Authentication and Authorization:** Enforcing robust authentication mechanisms, such as multi-factor authentication (MFA), to verify user identities. Authorization policies should define user permissions based on roles and responsibilities (Hu et al., 2006).
- **Role-Based Access Control (RBAC):** Implementing RBAC to ensure that users have access only to the resources necessary for their roles. RBAC helps minimize the risk of unauthorized access and data breaches (Sandhu et al., 1996).
- **Audit Logs:** Maintaining detailed audit logs of all access and activity related to AI systems. Logs should capture user actions, access attempts, and changes to data or configurations. Regularly reviewing audit logs helps detect and investigate suspicious activities (Kent & Souppaya, 2006).
- **Access Reviews:** Conducting periodic reviews of access controls and permissions to ensure they are up-to-date and aligned with security policies. Access reviews help identify and revoke unnecessary or outdated access rights (Furnell & Tsaganidi, 2004).

5.6. Bias Mitigation

Continuously monitoring and mitigating biases in AI models is essential to ensure fairness and accuracy:

- **Bias Detection:** Implementing tools and techniques to detect biases in AI models. Methods such as fairness-aware machine learning and statistical tests can identify disparities in model outcomes based on demographic groups (Mehrabi et al., 2021).
- **Bias Mitigation Techniques:** Applying techniques such as reweighting, resampling, and adversarial debiasing to reduce biases in training data and model outputs. These methods help create more balanced and fair AI systems (Bellamy et al., 2019).
- **Diverse Datasets:** Ensuring that training datasets are representative of diverse populations and scenarios. Diverse datasets help reduce biases and improve the generalizability of AI models (Buolamwini & Gebru, 2018).
- **Fairness Audits:** Conducting regular fairness audits to evaluate the impact of AI models on different demographic groups. Fairness audits involve assessing model performance, outcomes, and potential biases, and making necessary adjustments (Raji et al., 2020).

5.7. Software and Hardware Security

Ensuring that the software and hardware components of AI systems are secure and regularly updated to address known vulnerabilities is vital:

- **Secure Software Development Lifecycle (SDLC):** Adopting secure SDLC practices to integrate security into every phase of software development. This includes threat modeling, secure coding practices, code reviews, and security testing (NIST, 2004).
- **Vulnerability Management:** Regularly scanning for and addressing vulnerabilities in software and hardware components. This involves applying patches, updates, and security fixes promptly to mitigate risks (Scarfone & Mell, 2007).
- **Secure APIs:** Ensuring that APIs used by AI systems are secure and follow best practices. This includes implementing authentication, authorization, encryption, and input validation to prevent API-related vulnerabilities (Richardson & Ruby, 2007).
- **Hardware Security:** Utilizing secure hardware platforms that provide built-in security features, such as Trusted Platform Modules (TPMs) and secure enclaves. Hardware security helps protect AI systems from physical tampering and side-channel attacks (Gupta & Kim, 2020).
- **Security Testing:** Conducting thorough security testing, including static analysis, dynamic analysis, and fuzz testing, to identify and mitigate software vulnerabilities. Security testing helps ensure that AI systems are resilient against attacks (McGraw, 2004).

VI. Limitations

Despite the comprehensive approach of this study, several limitations constrain the scope and applicability of its findings see figure 2 below.. Understanding these limitations is crucial for interpreting the results and guiding future research efforts.

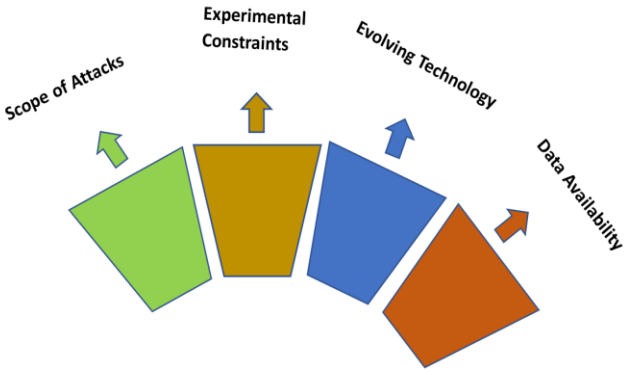

Figure 2: Limitations of AI

6.1. Scope of Attacks

One significant limitation is the focus on known attack vectors. While this study provides an in-depth analysis of well-documented vulnerabilities such as data poisoning, adversarial attacks, model inversion, and model stealing, it may not fully address emerging threats.

- **Rapid Evolution of Attack Techniques:** Cybersecurity threats evolve rapidly, and attackers continuously develop new techniques to exploit vulnerabilities in AI systems. This dynamic nature means that the study might not cover the latest or the most sophisticated attack methods that have not yet been documented in the literature (Sommer & Paxson, 2010).
- **Lack of Comprehensive Taxonomy:** The diversity and complexity of AI and cybersecurity landscapes make it challenging to develop a comprehensive taxonomy of all possible attack vectors. Consequently, some attack methods might be overlooked or insufficiently explored, leading to an incomplete threat model (Szegedy et al., 2014).
- **Focus on Commonly Studied Attacks:** The study predominantly examines attacks that have been extensively studied and reported in academic literature. This focus can inadvertently neglect less common or novel attacks that could be equally or more harmful. Emerging threats such as AI model watermarking attacks or side-channel attacks on AI hardware are examples of areas that require further exploration (Guo et al., 2018).

6.2. Experimental Constraints

Controlled experiments, while essential for understanding specific vulnerabilities, may not fully capture the complexity and variability of real-world environments.

- **Simplified Experimental Conditions:** In experimental settings, researchers often simplify conditions to isolate specific variables and better understand their effects. However, these simplifications can lead to findings that do not fully translate to more complex real-world scenarios (Hutson, 2017). For instance, laboratory conditions may not account for the diverse range of inputs and environmental factors that an AI system would encounter in a real-world deployment.
- **Limited Scope of Simulations:** Simulations used in experiments may not encompass the full range of potential threats or system behaviors. The controlled nature of these experiments means that certain interactions and interdependencies present in operational environments are not adequately represented (Baumann et al., 2021).
- **Replicability and Scalability Issues**: Experimental studies often struggle with replicability and scalability. The results obtained in a controlled setting might not be replicable in a different context or at a larger scale, limiting the generalizability of the findings (Peng et al., 2011).

6.3. Evolving Technology

The rapidly evolving nature of AI and cybersecurity technologies poses significant challenges for maintaining the relevance and accuracy of the study's findings.

- **Technological Advancements:** AI and cybersecurity technologies are advancing at a fast pace, with new algorithms, tools, and frameworks being developed continuously. These advancements can render current findings obsolete or less applicable over time (LeCun et al., 2015). For example, new machine learning techniques like federated learning introduce new paradigms and potential vulnerabilities that were not previously considered.

- **Lag Between Research and Practice:** There is often a time lag between when new technologies are developed and when they are widely adopted in practice. This lag can result in a discrepancy between the state-of-the-art in research and what is implemented in real-world systems (Zhu et al., 2019). Consequently, the study's recommendations may be based on technologies that are not yet widely deployed or may become outdated as new solutions are introduced.
- **Changing Threat Landscape:** As both AI and cybersecurity evolve, so does the threat landscape. New vulnerabilities and attack vectors continuously emerge, requiring ongoing research and adaptation of security measures. Keeping up with these changes is a perpetual challenge, making it difficult to provide definitive solutions (Papernot et al., 2018).

6.4. Data Availability

Access to detailed data on real-world incidents is limited due to the sensitive nature of cybersecurity breaches, posing a significant challenge for comprehensive analysis.

- **Confidentiality and Privacy Concerns:** Organizations are often reluctant to share detailed information about security incidents due to confidentiality and privacy concerns. This reluctance can limit the availability of data for research purposes, hindering the ability to analyze real-world attack patterns and vulnerabilities (Mell & Grance, 2011).
- **Incomplete Data Sets:** Even when data is available, it may be incomplete or lack critical details necessary for thorough analysis. Missing information can skew the results and limit the study's ability to draw accurate conclusions about the nature and impact of specific vulnerabilities (Rahm & Do, 2000).
- **Bias in Available Data:** The data that is available might be biased towards certain types of incidents or industries. For example, high-profile breaches in large organizations may be more frequently reported and studied than smaller-scale incidents affecting small and medium-sized enterprises (SMEs). This bias can lead to an incomplete understanding of the broader threat landscape (Lin, 2016).
- **Regulatory and Legal Barriers:** Regulatory and legal barriers can also restrict access to incident data. Compliance with data protection regulations, such as the General Data Protection Regulation (GDPR) in the European Union, imposes strict controls on data sharing and usage, complicating research efforts (Voigt & Von dem Bussche, 2017).

VII. Conclusion

The integration of AI into cybersecurity systems introduces new vulnerabilities that must be addressed to ensure robust defense mechanisms. By understanding and mitigating threats such as data poisoning, adversarial attacks, model inversion, and others, we can enhance the security and reliability of AI-driven cybersecurity tools. These case studies collectively illustrate the diverse range of vulnerabilities that AI systems can face when integrated into cybersecurity contexts. From adversarial attacks and data poisoning to model inversion and AI-enhanced malware, the threats are multifaceted and evolving. Addressing these vulnerabilities requires a combination of robust training practices, continuous monitoring, rigorous validation, and advanced privacy-preserving techniques. By learning from real-world incidents and proactively developing mitigation strategies, we can better safeguard AI-integrated systems against the complex landscape of cyber threats.

The limitations outlined above highlight the challenges inherent in studying the vulnerabilities introduced by the integration of AI into cybersecurity systems. While the study provides valuable insights and recommendations, it is essential to recognize the constraints imposed by the scope of attacks, experimental conditions, the evolving nature of technology, and data availability. Future research should aim to address these limitations by expanding the scope of studied attacks, improving the realism of experimental settings, staying abreast of technological advancements, and advocating for better data sharing practices. By doing so, we can develop more comprehensive and resilient strategies to protect AI-driven cybersecurity systems. Future research should focus on developing more resilient AI models and exploring new defensive strategies to keep pace with evolving threats.

VIII. References

1. Aggarwal, C. C., & Yu, P. S. (2008). A general survey of privacy-preserving data mining models and algorithms. In Privacy-preserving data mining (pp. 11-52). Springer.

2. Antunes, N., & Vieira, M. (2015). Defending against web application vulnerabilities. Computer, 48(6), 53-61.
3. Baumann, R., Wettig, D., Moradian, E., & Sørensen, L. (2021). Bridging the gap between academic research and real-world cyber threat intelligence. Computers & Security, 107, 102310.
4. Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. (2019). AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. IBM Journal of Research and Development, 63(4/5), 4-1.
5. Biggio, B., Nelson, B., & Laskov, P. (2012). Poisoning attacks against support vector machines. Proceedings of the 29th International Coference on International Conference on Machine Learning (pp. 1467-1474). Omnipress.
6. Breck, E., Polyzotis, N., Roy, S., Whang, S. E., & Zinkevich, M. (2019). Data validation for machine learning. In Proceedings of the 2nd International Workshop on Data Management for End-to-End Machine Learning (pp. 1-4).
7. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cybersecurity intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.
8. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91).
9. Cárdenas, A. A., Amin, S., Lin, Z. S., Huang, Y. L., Huang, C. Y., & Sastry, S. (2011). Attacks against process control systems: risk assessment, detection, and response. In Proceedings of the 6th ACM symposium on information, computer and communications security (pp. 355-366).
10. Dasu, T., & Johnson, T. (2003). Exploratory data mining and data cleaning. John Wiley & Sons.
11. Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322-1333).
12. Freire, J., Koop, D., Santos, E., & Silva, C. T. (2008). Provenance for computational tasks: A survey. Computing in Science & Engineering, 10(3), 11-21.
13. Furnell, S., & Tsaganidi, V. (2004). Security auditing—what, where, when and by whom
14. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).
15. Guo, Y., Zhang, L., Hu, J., Zhao, X., & Zhou, M. (2018). Automated unit test generation for AI-based applications. Proceedings of the 2018 International Conference on Software Engineering (ICSE), 13-23.
16. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. r., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82-97.
17. Hutson, M. (2017). Artificial intelligence faces reproducibility crisis. Science, 359(6377), 725-726.
18. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.
19. Lin, H. (2016). Cyber conflict and international relations. In J. Baylis, S. Smith, & P. Owens (Eds.), The globalization of world politics: An introduction to international relations (pp. 278-289). Oxford University Press.
20. Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. National Institute of Standards and Technology.
21. Neff, G., & Nagy, P. (2016). Automation, algorithms, and politics| talking to Bots: Symbiotic agency and the case of Tay. International Journal of Communication, 10, 17.
22. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). Towards the science of security and privacy in machine learning. Proceedings of the IEEE European Symposium on Security and Privacy (EuroS&P), 399-414.
23. Peng, R. D., Dominici, F., & Zeger, S. L. (2011). Reproducible epidemiologic research. American Journal of Epidemiology, 163(9), 783-789.
24. Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.
25. Richardson, L., & Ruby, S. (2007). RESTful web services. O'Reilly Media, Inc.
26. Scarfone, K., & Mell, P. (2007). Guide to intrusion detection and prevention systems (IDPS). NIST Special Publication, 800, 94.
27. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. Proceedings of the 2010 IEEE Symposium on Security and Privacy (SP), 305-316.

28. Stoecklin, M. P., Wang, H., & He, Y. (2018). DeepLocker: How AI can power a stealthy new breed of malware. IBM Research AI.
29. Strickland, E. (2019). IBM Watson, heal thyself: How IBM overpromised and underdelivered on AI health care. IEEE Spectrum. Retrieved from https://spectrum.ieee.org/ibm-watson-health
30. Stouffer, K., Falco, J., & Scarfone, K. (2011). Guide to industrial control systems (ICS) security. NIST Special Publication, 800, 82.
31. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. Proceedings of the International Conference on Learning Representations (ICLR).
32. Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. 25th USENIX Security Symposium (USENIX Security 16), 601-618.