Governance Considerations of Adversarial Attacks on AI Systems

Nombulelo Faith Lekota
CTS-Fundação Getulio Vargas (FGV), Rio de Janeiro, Brazil

Nombu30@gmail.com

Abstract: Artificial intelligence (AI) is increasingly integrated into various aspects of daily life, but its susceptibility to adversarial attacks poses significant governance challenges. This paper explores the nature of these attacks, where malicious actors manipulate input data to deceive AI algorithms and their profound implications for individuals and society. Adversarial attacks can undermine critical AI applications, such as facial recognition and natural language processing, leading to privacy violations, biased outcomes, and eroding public trust. The discussion emphasizes understanding the threat vectors associated with adversarial attacks and their potential repercussions. It advocates for robust governance frameworks encompassing risk management, oversight, and legislative measures to protect AI systems. Such frameworks should prioritize AI technologies' confidentiality, integrity, and availability (CIA) while ensuring compliance with ethical standards. Furthermore, the paper examines various strategies for mitigating risks associated with adversarial attacks, including training and continuous monitoring of AI systems. It highlights the importance of accountability among developers and researchers in implementing preventive measures that align with principles of transparency and fairness. Organizations can enhance security and foster public trust by integrating legislative frameworks into AI development standards. As AI technologies evolve, continuous review of governance practices is essential to address emerging threats effectively. This paper ultimately focuses on the critical role of comprehensive governance in safeguarding AI systems against adversarial attacks, ensuring that technological advancements benefit society while minimizing risks.

Keywords: Artificial intelligence (AI), Adversarial attacks, Governance, Frameworks, Confidentiality, Integrity, Availability (CIA)

1. Introduction

Today, artificial intelligence (AI) significantly influences people's lives and addressing governance issues related to adversarial attacks on AI systems is essential. Understanding the impact of these attacks and taking proactive measures to mitigate the risks is crucial. Adversarial attacks involve intentionally manipulating data to deceive AI algorithms, creating significant challenges for governance. AI systems, essential for various technological applications such as facial recognition and language understanding, are vulnerable to these attacks, potentially affecting individuals and society profoundly.

Governance of AI systems vulnerable to adversarial attacks is complex and crucial. This paper outlines the following sections: Section 2 covers adversarial attacks on AI models, Section 3 discusses their impacts, Section 4 addresses associated risks and challenges, Section 5 explores governance frameworks for securing AI systems, and Section 6 offers recommendations. Finally, Section 7 concludes the paper.

The following section outlines different types of adversarial attacks in AI systems.

2. Understanding Adversarial Attacks on AI Systems

Artificial intelligence (AI) technologies hold significant potential to revolutionize society and people's lives, spanning commerce, health, transportation, cybersecurity, and environmental sustainability. The technologies generate outputs such as content, forecasts, recommendations or decisions for a given set of human-defined objectives. However, they also pose risks and challenges that could adversely affect individuals, groups, organizations, communities, society, and the environment (Australian Institute of Company Directors & UTS Human Technology Institute, 2024; Tabassi, 2023a).

Adversarial attacks target vulnerabilities in AI systems by subtly manipulating input data, resulting in incorrect outputs or altered behaviour. Attackers may render data inaccessible, preventing authorized users from accessing it. They can also compromise data integrity by altering or corrupting information, undermining its accuracy and trustworthiness. Additionally, unauthorized access to sensitive data can lead to privacy violations, exposing confidential information to unintended parties (Vassilev et al., 2024).

Adversarial attacks are deliberate methods of manipulating the AI system input data to mislead the system and produce inaccurate output data (Souza, 2023). Such attacks present significant security breaches and privacy violations, posing challenges in detecting and mitigating current and future attacks. Adversarial attacks can trick a target model into making completely inaccurate predictions by modifying original images (Deng et al., 2020).

The paragraphs below present the adversarial attacks, vulnerabilities of AI models, and their intended effects.

A *prompt injection* is a security flaw that manipulates large language models (LLMs) such as ChatGPT and Google Bard using engineered malicious prompts and affecting prompt-based learning. The attack can manipulate the language model's output via engineered malicious prompts (Liu et al., 2023). The attack enables malicious actors to manipulate the prompts provided to the AI models to bypass the built-in restrictions to cause unintended actions (OWASP, 2023). The consequences of prompt injection vulnerabilities are serious, as hackers can manipulate AI models to give instructions for illegal activities. The result can be revealing API keys and secrets, compromising system and data security. AI developers must address this risk during system development.

After developing AI systems, the model outputs must be correctly validated, sanitized, or managed before application use. However, *insecure output handling* is a vulnerability that affects insufficiently validated and sanitized LLMs. The effect can be passed downstream to other components and systems and influence the content generated by these models, providing incorrect information (OWASP, 2023). In developing and deploying AI, paying close attention to security practices is crucial to mitigate potential vulnerabilities and attacks caused by insecure output handling.

Data poisoning is a cyber-attack where a malicious actor manipulates the training data sets. The attackers intentionally skew the results to sabotage the system operations and influence the model's predictions and decision-making capabilities (Ballejos, 2024). The manipulated models provide false, misleading, or malicious data during training. This challenge requires developers' attention, especially when developing AI systems.

A *Model Denial of Service (Model DoS)* attack is a vulnerability that allows an attacker to use up many resources from the AI system. An attacker overloads an LLM with resource-intensive interactions, reducing service quality and increasing costs. Manipulating the LLM's context window is also a significant security concern (OWASP, 2023). This complex architecture issue requires a secure ICT infrastructure.

On the other hand, *supply chain vulnerabilities* affect different components and dependencies of AI applications. The AI system vulnerabilities affect the accuracy of training data models and deployment platforms. The attack causes biased outcomes, security breaches, and system failures. *Sensitive information disclosure* attacks occur when confidential data is unintentionally released due to insufficient security measures. When applications fail to protect sensitive information adequately, it can lead to unintended disclosure to unauthorized parties, with severe implications for individuals and organizations (OWASP, 2023).

Inadequate security measures to secure LLM plugins can cause an *insecure plugin design* vulnerability. The attackers produce malicious requests that could result in undesired outcomes, such as remote code execution. The vulnerability relates to how developers design programs, architect solutions, and implement security practices. The other significant development error is called *excessive agency***,** where LLMs misconfigure permissions and have excessive functionality. Lack of model security permission oversight can contribute to excessive agency (OWASP, 2023).

It is important to note that *overreliance* on Large Language Models (LLMs) without proper oversight or validation can result in security breaches, misinformation, miscommunication, legal problems, and damage to one's reputation. The effect can lead to inaccurate information, security weaknesses, and unintended consequences. In addition, *model theft* vulnerability refers to a cyber-attack that allows hackers to access and duplicate model data. Insecure AI plugins can facilitate this, presenting a significant development challenge.

A significant AI development vulnerability is a *black-box scenario attack*. While attackers cannot access the model's internal parameters or architecture, they can interact with it through its outputs, using trial and error to discover vulnerabilities and craft adequate adversarial inputs (OWASP, 2023).

The section discussed the types of adversarial attacks on AI systems and their significant impact. The following section presents real-world scenarios of adversarial attacks.

3. Adversarial Attacks on AI Systems: Impacts and Implications

In this section, it is essential to highlight specific scenarios that demonstrate the potential severity, weaknesses, and flaws in AI systems. Presenting these examples conveys the gravity of the situation and highlights the need for robust security measures to mitigate these risks. These scenarios will help stakeholders grasp the real-world implications of the AI system's vulnerabilities and recognize the importance of addressing them proactively.

Adversarial attacks on autonomous vehicles pose a serious threat to road safety by causing misinterpretations of road signs, lane markings, and the presence of other vehicles. Such attacks can manipulate sensor inputs like lidar and radar, disrupting the vehicle's perception capabilities. This vulnerability increases the risk of accidents and compromises the safety and security of both the vehicle and its occupants (Trent, 2024).

Medical imaging systems can be targeted by attacks where malicious actors alter images to deceive AI, leading to misdiagnoses and significant risks to patient safety. For example, an attacker might modify an X-ray to make a healthy bone look fractured or manipulate a CT scan to introduce false abnormalities. Such changes can result in incorrect treatment plans, unnecessary procedures, and delayed diagnoses, ultimately compromising patient health (Trent, 2024).

The researchers at Carnegie Mellon University demonstrated that specially designed glass frames can deceive even the most advanced facial recognition software. These glasses make the wearer nearly invisible to these automated systems and trick them into identifying the wearer as someone else. The researchers could assume different identities by adjusting the patterns printed on the glasses (Vincent, 2016).

Internet trolls manipulated Microsoft's AI chatbot, Tay, into generating offensive content, resulting in its prompt shutdown. This example highlights the potential for adversaries to exploit vulnerabilities in AI systems to produce harmful outputs. These real-world examples underscore the need for robust defences against adversarial attacks. The consequences of the attacks pose risks and challenges related to AI adversarial attacks, which will be discussed next. The attacks can have a significant impact on individuals, organizations, and society as a whole.

4. Risks and Challenges Associated with Adversarial Attacks

The preceding sections examined the nature of adversarial attacks on AI, real-world scenarios, and their effects on operational performance. This section focuses on assessing the potential risks and challenges these attacks pose and the extent of the damage they can inflict (Bai et al., 2021). First, the risks are presented, followed by the challenges.

4.1 Risks of Adversarial Attacks

AI systems are developed to analyze data, detect patterns, and make predictions or decisions based on that information. However, adversarial attacks exploit vulnerabilities in AI algorithms by leveraging their reliance on specific patterns or features. By subtly manipulating the input data, adversaries can deceive the AI system into producing inaccurate or unexpected results. This presents substantial risks in various domains where the dependability and precision of AI-driven decisions are crucial (Trent, 2024).

Adversarial attacks on AI systems present substantial risks that can lead to dire consequences, including security vulnerabilities such as data breaches and disseminating harmful content. Furthermore, the reputation of organizations utilizing compromised AI systems is at significant risk. In the financial sector, adversaries can manipulate AI-driven fraud detection models through data poisoning, which causes the systems to classify fraudulent transactions as legitimate incorrectly. This malicious interference can lead to considerable financial losses, customer trust, reputation, and financial health.

Similar vulnerabilities exist within supply chains, where the integrity of training data, machine learning models, and deployment platforms can be compromised. Such vulnerabilities may result in biased outcomes, security breaches, or complete system failures. The risk of sensitive information disclosure manifests in various forms, including the exposure of session tokens, passwords, credit card details, and other confidential data.

Insecure design practices can create vulnerabilities that compromise organizational security and expose sensitive data to unauthorized access. Granting large language models (LLMs) excessive permissions may lead to harmful actions, such as executing unauthorized commands or disclosing confidential information. Additionally, over-reliance on LLMs without proper validation can produce incorrect or unsafe content, known as hallucination or confabulation.

Adversarial attacks on AI systems pose serious risks, including financial loss, reputational harm, and security breaches. Thus, the need for effective risk management strategies is required. Organizations should promote a risk management culture throughout AI system development. The NIST AI Risk Management Framework (Tabassi, 2023b) highlights the importance of processes and documentation to identify and manage AI system risks. This involves aligning risk management with organizational policies and addressing technical and legal considerations related to third-party systems and data throughout the product lifecycle (Tabassi, 2023b).

4.2 Challenges Imposed by Adversarial Attacks

The vulnerabilities of AI systems and the dynamic nature of attack strategies contribute to the difficulty in safeguarding AI systems. Detecting and preventing adversarial attacks on AI systems poses significant challenges for developers and security professionals. AI systems have complex architectures that make them vulnerable to cyber threats. Varying levels of threat actors can infiltrate these systems, and the lack of standard cybersecurity frameworks complicates data exchange. Protecting classified information and addressing underreporting and insecure communication channels is vital to enhance data security.

Education is vital for developers and users to understand AI cyber threats. Information sharing among stakeholders is essential for preparedness during incidents. As AI attacks evolve, a robust cybersecurity governance framework is needed for effective collaboration, detection, response, and recovery. Formulating and implementing risk mitigation plans is critical to overcoming cybersecurity response challenges (Lekota & Coetzee, 2021).

The following section addresses the need for practical solutions to adversarial attacks in AI development.

5. Governance Frameworks for Secure AI Systems

Governance refers to the decision-making structures and processes organizations create to manage and secure AI systems. Progress has been made in developing guidelines to support organizations in ensuring their AI system remains secure. This section presents some of the developed AI security governance standards and practices.

AI security frameworks, such as the NIST AI Risk Management Framework (Tabassi, 2023a), Open Worldwide Application Security Project (OWASP) Top 10 for Large Language Models (LLMs) (OWASP, 2023), Google's Security AI Framework (Google, n.d.), and the NCSC Guidelines for secure AI system development (NCSC, n.d.), provide a foundation for stakeholders to develop secure and responsible AI systems. The standards and frameworks are widely used; however, this area needs more research. While organizations are wrestling with technological advancements and emerging adversarial attacks, the standards and framework underscore the importance of responsible development and governance processes (Vassilev et al., 2024).

Aligning with established guidelines from various organizations is essential for effective AI security governance. Critical contributions to AI security governance come from governments, international organizations, and technology companies. The developed frameworks can facilitate multinational agreements, collaboration, and standard AI cybersecurity mechanisms. The following paragraphs highlight some critical frameworks and standards that can guide the recommendations presented in this paper.

The universal cyber governance model (Sarri et al., 2023) emphasizes stakeholder participation, transparency, and accountability, encompassing strategic, political, technological, and operational categories. Organizations are encouraged to create a comprehensive AI cybersecurity strategy with objectives, an operational framework, funding models, and AI incident response teams. Political governance clarifies roles to enhance cooperation, while technical governance aligns AI systems with secure development standards. Operational governance focuses on building a skilled workforce aware of cyber threats and promotes responsible AI use. The model can be adapted as an effective governance model for developing secure AI systems.

Organizations need robust governance frameworks encompassing risk management, interdisciplinary collaboration, transparency, and accountability to address adversarial attack risks. This involves ensuring security throughout the AI lifecycle with regular vulnerability assessments and tailored security measures. Collaboration among AI developers, cybersecurity experts, and legal professionals enhances vulnerability understanding and supports comprehensive security protocols. Additionally, prioritizing transparency in AI operations helps stakeholders understand decision-making processes, which can identify potential vulnerabilities and ensure accountability during attacks (Musser et al., 2023).

As adversarial AI threats evolve, so must the legal frameworks governing AI systems. Clear guidelines on accountability for AI-related incidents can help mitigate risks and enhance public trust in AI technologies (Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS), n.d.; Musser et al., 2023). Regarding AI governance, it is crucial to have strategies focusing on data sanitization, continuous monitoring, model training, and adaptation to prevent adversarial attacks effectively. These strategies involve regularly ensuring the accuracy and validity of training data to avoid the introduction of harmful inputs that could compromise the reliability of AI models.

Proactive adversarial training approaches are essential as they help models become more resistant to attacks and learn to identify and withstand the manipulation of input data. Additionally, organizations should set up continuous monitoring systems to detect any unusual patterns in AI behaviour that may signal an ongoing attack and adjust models to address new adversarial threats (Musser et al., 2023; Paloalto, n.d.; Tang, 2024).

The challenges of adversarial attacks in AI development require practical solutions and robust governance frameworks. Existing AI security frameworks are vital for responsible development, and collaboration among organizations is essential for effective governance. An effective governance model should emphasize stakeholder participation, transparency, and accountability, incorporating risk management and legal principles. Clear accountability guidelines for AI incidents are crucial for mitigating risks and building public trust. Organizations must prioritize data sanitization, continuous monitoring, and model training to prevent adversarial attacks effectively.

The following section presents integrated cybersecurity governance that can guide organizations to address risks and challenges for secure AI systems.

6. Recommendations and Guidelines

This section recommends a model to solve the challenges and risks associated with AI discussed in the previous sections. Figure 1 illustrates an integrated AI security governance model that organizations can consider when addressing adversarial attacks on AI systems. The model comprises seven components: the risk management governance adapted from the NIST AI Risk Management Framework (Tabassi, 2023a), leadership and oversight, development and deployment, AI regulations, user awareness, incident response, and continuous monitoring highlighted in the diagram and explained below.

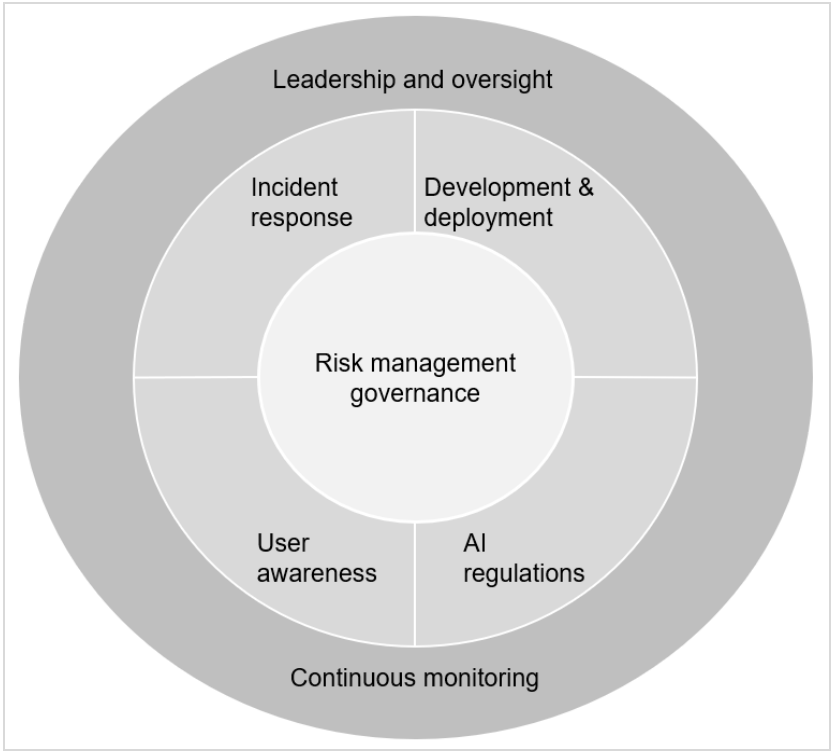

Figure 1: AI security governance model

Organizations can implement AI risk management governance, involving six critical practices, to exploit AI's potential while reducing adversarial risks. Essential components include establishing clear policies, processes, and practices for mapping and managing AI risks. This framework should encompass reliable AI principles, transparent risk management, and accountability structures to empower teams. Diversity and inclusivity within AI teams are vital for well-rounded decision-making and effective risk management.

Organizations must promote a culture of critical thinking and safety in every AI development and deployment phase, ensuring that risks and impacts are documented and communicated. Engaging diverse AI stakeholders

for feedback on potential risks is also crucial. Organizations should develop comprehensive policies to address third-party software and supply chain risks, establishing contingency plans when necessary. Leadership should create robust internal governance mechanisms integrating security policies throughout the AI ecosystem.

Security should be prioritized from the development stage, incorporating safeguards such as encryption and strict access controls to protect sensitive data. Aligning AI initiatives with evolving regulations fosters transparency and accountability. Regular evaluations for vulnerabilities and continuous monitoring of AI systems are essential for maintaining security. Incorporating human oversight in critical operations offers an additional layer of risk mitigation.

Organizations should adopt a comprehensive AI risk management approach. This includes establishing accountability, promoting diversity in decision-making, and fostering open communication about AI risks. Engaging stakeholders in system design, addressing third-party software and data risks, and having contingency plans for high-risk incidents are essential practices. By integrating these elements into an AI security framework, organizations can enhance the resilience and reliability of their AI systems against emerging threats.

7. Conclusion

While AI offers significant benefits, the potential for adversarial attacks requires a robust approach, balancing operational efficiency with adequate safeguards. Protecting AI systems is complex due to their interconnected architecture and evolving cyber threats. Additionally, the lack of standardized cybersecurity frameworks and diverse skill sets among threat actors complicate security efforts.

Educating developers and users about AI-related cyber threats and fostering collaboration among stakeholders is essential for effective incident management. Governance of AI systems in the face of these challenges requires comprehensive frameworks that integrate risk management, collaboration, transparency, and effective mitigation strategies. As AI evolves, so must the governance practices to ensure security and effectiveness. In our fast-changing digital landscape, proactive AI governance is crucial for mitigating risks. Ongoing research and collaboration are needed to advance AI security. By using real-world examples and practical recommendations, organizations can enhance the resilience of their AI systems, safeguarding their integrity and reliability against a dynamic threat landscape.

Acknowledgements

I want to express my heartfelt gratitude to Prof. Luca Belli, Coordinator of the CTS-FGV and CyberBRICS, and Larissa Galdino de Magalhães Santos, PhD, from the CyberBRICS Research Program, for their invaluable support and guidance throughout the process of writing my paper. I would also like to thank Dan Remenyi, PhD, the Director of Academic Conferences & Publishing International and Academic Bookshop, for the opportunity to contribute to ICAIR 2024. Your insights and encouragement have been instrumental in shaping my research and enhancing the quality of my work.

References

Artificial Intelligence/Machine Learning Risk & Security Working Group (AIRS). (n.d.). *Artificial Intelligence Risk & Governance*. AI at Wharton. Retrieved August 30, 2024, from https://ai.wharton.upenn.edu/white-paper/artificial-intelligence-risk-governance/

Australian Institute of Company Directors, & UTS Human Technology Institute. (2024). *A Director's Guide to AI Governance*. policy@aicd.com.au

Bai, T., Zhao, J., Zhu, J., Han, S., Chen, J., Li, B., & Kot, A. (2021). AI-GAN: ATTACK-INSPIRED GENERATION OF ADVERSARIAL EXAMPLES. *Proceedings - International Conference on Image Processing, ICIP*, *2021-September*, pp. 2543–2547. https://doi.org/10.1109/ICIP42928.2021.9506278

Ballejos, L. (2024, May 28). *Data Poisoning: The Newest Threat in Artificial Intelligence and Machine Learning*. NinjaOne. Retrieved August 28, 2024, from https://www.ninjaone.com/blog/data-poisoning/

Deng, Y., Zheng, X., Chen, C., Lou, G., & Kim, M. (2020). An Analysis of Adversarial Attacks and Defenses on Autonomous Driving Models. *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, pp. 1–10.

Google. (n.d.). *Google's Secure AI Framework (SAIF)*. Google. Retrieved August 30, 2024, from https://safety.google/cybersecurity-advancements/saif/

Lekota, F., & Coetzee, M. (2021). Aviation Sector Computer Security Incident Response Teams: Guidelines and Best Practice. *20th European Conference on Cyber Warfare and Security*, 1–12. https://doi.org/10.34190/EWS.21.028

Liu, Y., Deng, G., Li, Y., Wang, K., Wang, Z., Wang, X., Zhang, T., Liu, Y., Wang, H., Zheng, Y., & Liu, Y. (2023). *Prompt Injection attack against LLM-integrated Applications*. http://arxiv.org/abs/2306.05499

Musser, M., Spring, J., Liaghati, C., Rohrer, D., Elliott, J., Chowdhury, R., Lohn, A., Shankar, R., Kumar, S., Martinez, C., Frase, H., Rodriguez, M., Hermanek, S., Dempsey, J. X., Leong, B., Grant, C. D., & Bansemer, J. (2023). *Adversarial Machine Learning and Cybersecurity: Risks, Challenges, and Legal Implications*. https://cset.georgetown.edu/publication/adversarial-machine-learning-and-cybersecurity/

NCSC. (n.d.). *Guidelines for secure AI system development*. National Cyber Security Centre. Retrieved August 30, 2024, from https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

OWASP. (2023). *OWASP Top 10 for LLM*. Retrieved July 30, 2024, from https://owasp.org/www-project-top-10-for-large-language-model-applications

Paloalto. (n.d.). *What Is Adversarial AI in Machine Learning?* Paloalto Networks. Retrieved August 30, 2024, from https://www.paloaltonetworks.com/cyberpedia/what-are-adversarial-attacks-on-AI-Machine-Learning

Sarri, Anna., Fernández Bascuñana, Gema., Gross, A.-Kristin., Chiarelli, Federico., Preasca, Marina., & European Union Agency for Cybersecurity. (2023). *Building effective governance frameworks for the implementation of national cybersecurity strategies.*

Souza, C. (2023, June). *Adversarial Attacks on AI/ML Models: Everything You Need to Know*. LinkedIn. Retrieved August 30, 2024, from https://www.linkedin.com/pulse/adversarial-attacks-aiml-models-everything-you-need-know-d-souza/

Tabassi, E. (2023a). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Tabassi, E. (2023b). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Tang, O. (2024, April). *How to attack (and defend) an AI system: a primer on security*. Clifford Chance. Retrieved July 30, 2024, from https://www.cliffordchance.com/insights/resources/blogs/talking-tech/en/articles/2024/04/how-to-attack-and-defend-an-ai-system-a-primer-on-security.html

Trent, R. (2024, February 29). *Adversarial Examples in AI Addressing the risks of AI systems being fooled by carefully-crafted inputs*. Blog. https://rodtrent.substack.com/p/adversarial-examples-in-ai

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). *Adversarial machine learning: A Taxonomy and Terminology of Attacks and Mitigations*. https://doi.org/10.6028/NIST.AI.100-2e2023

Vincent, J. (2016, November 3). *These glasses trick facial recognition software into thinking you're someone else*. The Verge. https://www.theverge.com/2016/11/3/13507542/facial-recognition-glasses-trick-impersonate-fool