6. Managing cyber risks in the face of AI- and ML - Driven Adversarial Attacks

Author: Godwill Chimamiwa

https://doi.org/10.70301/CONF.SBS-JABR.2024.1/1.6

Abstract

This paper presents a critical analysis of current cyber risk management practices in light of new and evolving Artificial Intelligence (AI) and Machine Learning driven adversarial attacks. Many enterprises are constantly grappling with cybersecurity risks and increased threats from phishing, ransomware and many other forms of cyber-attacks, often resulting in substantial financial losses when the risks are not adequately addressed. With the advent of Artificial Intelligence (AI) and Machine Learning (ML), such cyber-attacks and incidents will become more prevalent and potentially more devastating to businesses large and small. With AI and ML tools at their disposal, cybercriminals can dramatically reduce technical barriers for launching cyberattacks. They can easily develop more sophisticated social engineering tactics and 'deep fakes' that are not easily identifiable as such, thereby increasing the risks of unauthorized data disclosure. Drawing on literature review analysis, this research explores current and emerging AI- and ML-driven cyber threats faced by enterprises, effectiveness of current cyber mitigation measures and future management practices that can be leveraged to improve the security posture of enterprises. The study evaluates both technical and non-technical cyber risk management and mitigation measures and frameworks. The findings from this study help inform enterprise cyber risk managers and practitioners about the enormity of AI- and ML-driven cyber risks and presents emerging best practices to adequately mitigate those risks. The study contributes to the growing research on how threat actors are leveraging and AI and ML to expand cyber threats and how enterprises and organizations should respond to these ever evolving cyber risks.

Keywords: *cyber risk management, AI-driven, ML-driven, adversarial attacks, cyber risk frameworks*

INTRODUCTION

The rapid advancement and ubiquity of modern technologies such as Artificial Intelligence (AI) and Machine Learning (ML) has introduced significant opportunities but also substantial challenges in the cybersecurity threat landscape. Cyber threat actors are taking advantage of these technologies to create new and more potent adversarial attack vectors, threatening business operations and in some cases the very existence of ill-prepared and ill-equipped businesses. With generative AI and ML tools at their disposal, cybercriminals can dramatically reduce technical barriers for launching cyberattacks. They can easily develop more sophisticated social engineering tactics and 'deep fakes' that are not easily identifiable as such, thereby increasing the risks of unauthorized data disclosure. Cyber risk management is "a multifaceted approach aimed at identifying, evaluating, and mitigating the potential risks posed by cyber threats to an organization's digital assets, sensitive data, and critical infrastructure." (Mizrak, 2023). Identification, evaluation and mitigation

of legacy cyber risks is substantially different from AI- and ML-driven cyber risks as these can easily morph and therefore become more evasive to conventional cybersecurity mechanisms. To adequately address the challenges posed by AI- and ML-driven cyber threats, existing cyber risk management frameworks and approaches need to be revisited and new and more effective AI-aware mitigation measures and frameworks established. From an academic perspective, cyber risk management research in the context of AI and ML is still emerging and much more still needs to be done. New research findings also need to be converted to proven cyber risk management strategies and practices. Drawing on literature review analysis, this research explores current and emerging AI- and ML-driven cyber threats faced by enterprises, effectiveness of current cyber mitigation measures and future management practices that can be leveraged to improve the security posture of enterprises. The study evaluates both technical and non-technical cyber risk management and mitigation measures and frameworks. The findings from this study help inform enterprise cyber risk managers and practitioners about the enormity of AI- and ML-driven cyber risks and presents emerging AI-aware best practices to adequately mitigate those risks.

LITERATURE REVIEW

Many researchers focus on the positive uses of AI and not so much on adversarial AI-driven cyber risks. According to Kaloudi and Li (2020), "researchers have not summarized AI-based cyber attacks enough to be able to understand the adversary's actions and to develop proper defenses against such attacks." Similarly, Schreiber and Schreiber (2024) note that there is "insufficient exploration of AI cybersecurity awareness in the current literature". However, the research interest and focus is steadily growing. A number of interesting and relevant research papers on AI-driven cyber risks have emerged in the past five years, highlighting the increasing attention to this critical area. In their research paper titled "Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI", Malatji & Tolah (2024) delve into "the multifaceted dimensions of AI-driven cyberattacks" and provide some interesting insights on motivations, implications and potential mitigations. Wang et al. (2023) offer an overview of recent advancements in AI- and ML-adversarial attacks and defenses, with a special focus on machine learning and deep neural network-based classification model. Another recent research is from Kumar (2024), whose survey presents "more than 200 recent papers concerning adversarial attacks and techniques". From these and other research papers, interesting patterns on AI- and ML-driven cyber risks begin to emerge. Characteristics of these risks are different from the legacy cyber risks. For instance, AI and ML based cyber threats have a high level of sophistication and evasiveness (Malatji M. & Tolah A., 2024). AI-driven cyberattacks "involve using advanced machine learning algorithms to identify vulnerabilities, predict patterns and exploit weaknesses" (Dewayne, 2024). Other AI- and ML-driven cyber risks include "Deep Fakes" resulting "from the "democratization" of powerful generative AI technologies" (Lyu, 2024), "next-generation" and "AI-enhanced malware" (Fritsch et. al.) and "AI-powered tools that use data analysis for offensive cyber operations" (Yamin et. al., 2021). In addition, some researchers establish

that "self-learning" and "AI-powered" social and intelligent bots can potentially "unleash incredibly powerful, human-like armies of social bots, in potentially well coordinated campaigns of deception and influence" (Foysal et. al., 2019, Guembe et. al.). More complicated cases of AI-related cyber risks emanate from adversarial machine learning (ML), where a bad actor uses ML to alter the functionality of AI-designed models, resulting in an undesirable behaviour of the models (Mirsky et. al., 2022, Waizel, 2024).

EFFICACY OF CURRENT CYBER RISK MITIGATION PRACTICES

The efficacy of current cyber risk management practices in light of AI- and ML-driven adversarial attacks is a growing area of concern for practitioners. While traditional cyber risk management measures still play an important role in identifying, remediating and protecting critical enterprise infrastructure and assets, the sophistication and adaptability of AI and ML-driven attacks pose significant challenges. Many practitioners have primarily responded to the escalating cyber risk challenges by adopting technology solutions that automatically identify and block threats before they cause harm. With the advent of generative AI and ML, the best technical solutions become vulnerable as these can be easily circumvented as AI- and ML-tools are able to "adapt and change their attack method" on the fly (Hart, 2023). Cyber threat actors are leveraging AI and ML techniques such as AI-generated audio and video "deep fakes" and AI-enabled targeted and tailored phishing emails to overcome existing technical defenses. In 2020 cyber criminals used AI voice cloning to dupe a Hong Kong branch manager into authorizing $35 million in transfers, thinking he was acting on orders from his company's director. In 2018, cyber criminals managed to bypass a facial recognition authentication and authorization system using AI-generated synthetic images resembling authorized personnel. Practitioners need a more holistic approach, including 'human risk' management, collaboration on best practices, AI awareness and training, and AI governance, to adequately defend against cyber threats in the new era of AI- and ML-driven cyber attacks. Cyber risk management also needs to be repositioned within the enterprise to become a more encompassing domain across enterprise business lines. The role of the cyber risk manager or chief security officer is often marginalized to a mere IT function, devoid of influence in critical business decisions. Without a "seat at the table" when it comes to key business and AI transformation initiatives, cybersecurity practitioners will struggle to introduce and manage adequate security measures effectively, leaving their organization vulnerable to AI- and ML-driven cyber threats. Research studies show that the risks and trust issues posed by "black-box" AI systems are often neglected as companies seeking to leverage AI and ML for innovation, time to market and competitive advantage (Chakravorti, B., 2024, Carabantes, M., 2020). Cyber attackers can infiltrate and 'poison' data used for AI training models and manipulate the training model "to subvert the learning process" (Ishai et al., 2021). Early and proactive involvement and collaboration between the business and cyber risk managers is crucial to address 'potential' cyber risks that emanate from implementation of AI and ML initiatives (Aziz, S. & Dowling, M., 2019). This will ensure that these potential issues do not evolve to become vulnerabilities that can be exploited by threat actors down the line.

Another key question is to what extent existing Cyber Security Frameworks are capable of coping with AI- and ML-related cyber threats. Researchers and practitioners have begun to challenge the efficacy of existing cybersecurity frameworks in light of AI-driven cyber threats and propose new perspectives and improvements of existing frameworks. Malatji, M. & Tolah, A. (2024) propose a new "comprehensive framework for understanding adversarial and offensive AI", which offers new insights into potential, newer AI-aware mitigation strategies.

EMERGING CYBER RISK MITIGATION STRATEGIES AND BEST PRACTICES

Traditional cybersecurity measures remain important, but more emphasis should be put on emerging strategies and best practices that can deal more effectively with the complexity and sophistication of AI- and ML-driven cyber threats. Technical measures as well as other multi-pronged and more holistic approaches are important 'best practice' considerations.

Threat hunting

AI and ML capabilities are important tools for enhancing cybersecurity measures such as anomaly detection, malware identification and response, threat analysis and intrusion prevention, an approach termed 'threat hunting'. Using Natural Language Processing (NLP) and ML algorithms, AI and ML tools can sift through enormous amounts of data and identify patterns that indicate a potential manipulation or infiltration of the data by cyber criminals (Chaddad A. et. al., 2023). Instead of just reacting to cyber-attacks as and why they happen, this approach allows cyber risk practitioners to proactively address potential cyber threats before they can cause significant harm.

Adversarial training

Adversarial training is an approach whereby AI and ML defensive models are trained using 'adversarial' examples to test and improve their robustness. Such 'adversarial' examples consist of inputs designed to 'maliciously' manipulate the model in a safe and controlled environment. This is important approach, holding the promise to better understand how cyber threat actors leverage AI and ML to cause harm, which then helps develop more robust cybersecurity measures.

Research shows that adversarial training significantly enhances the effectiveness of cybersecurity systems. Research shows that adversarial training can improve the performance of cybersecurity systems (Nunez & Esteban, 2022). Alexander Shashkov et al. (2023) showed that attackers became more effective in "thwarting" cyber defenses when 'adversarial agent-learning' was used to optimize "adversarial behavior of agents cybersecurity simulations". These learnings and insights can be converted into powerful cyber defense mechanism as practitioners gain a more sophisticated understanding of potential uses of AL and ML by cyber threat actors.

Due to the black-box nature of AI and ML algorithms, there is always the residual risk that these algorithms might have been 'poisoned' and therefore subject to manipulation by threat actors.

Explainable AI

Explainable AI (XAI) refers to AI systems designed to provide clear, understandable explanations for their decisions and actions

(Minh et. al., 2022). This provides transparency and makes it easier to determine whether the AI and ML models might have been manipulated. The ability to explain AL and ML algorithms is a double-edged sword. "Current XAI models are still vulnerable to adversarial attacks, leading to public concerns about XAI security" (). Cyber threat actors can leverage the same technology to better understand defensive AI and ML models, making it easier for them to device infiltration measures. Anything technological in nature is bound to be an 'arms race' between practitioners and threat actors. These technical approaches must be combined with other non-technical measures to improve cybersecurity.

Adapted frameworks for AI and ML

Legacy cyber risk management frameworks do not adequately address emerging issues around AI- and ML-driven adversarial cyber threats. Practitioners and academic researchers have proposed new and more holistic frameworks that seek to address this gap.

The National Institute of Standards and Technology (NIST) has recently updated its cybersecurity framework to include AI and ML considerations (National Institute of Standards and Technology, 2024). to enhance threat detection and response capabilities. The International Organization for Standardization (ISO) also recently updated its standards to include guidelines for managing AI and ML cyber risks. These and other efforts represent important steps but they are just the first steps in a dynamic and rapidly evolving space. More attention needs to be given to improving existing frameworks to become more robust and responsive in the face of AI and ML risks.

In their thesis on "AI-driven cyber risk management framework", Agzayal & Bouhorma (2024) propose integrating AI with traditional risk management models to evolving cyber risks on smart cities and Internet of Things (IOT) ecosystems. Although this research is only specific to smart cities and IOT ecosystems, such efforts are important to advance legacy cyber risk management to cope with AL and ML-adversarial attacks. Similar efforts still need to be had, which this paper is one.

To be successful against cyber threat actors in the AI and ML era, cyber risk management frameworks must offer holistic models that not only focus on technology but address other non-technical aspects, including people, process and governance as shown in the figure below.

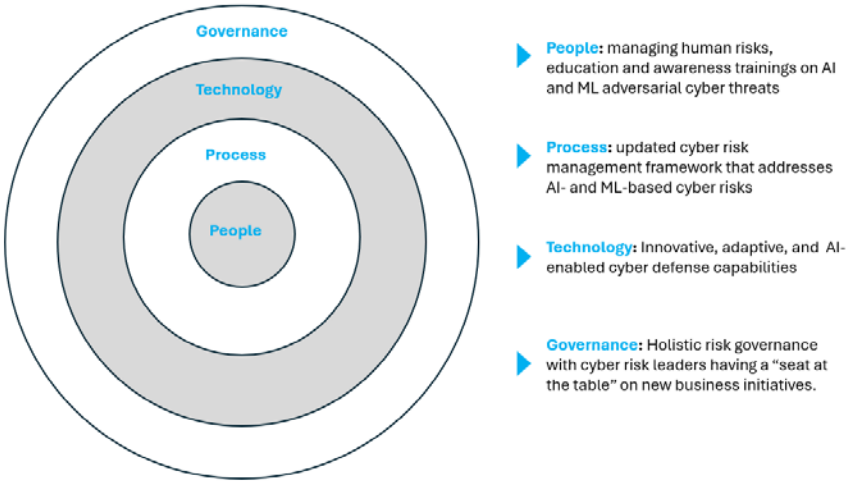

Figure 1 *– Holistic approach to managing cyber risks in the face to AI- and ML-driven cyber threats (author)*

In addition, it is important to constantly evaluate, measure and improve the maturity of an enterprise's cyber risk management practices in terms of its ability to cope with evolving AI and ML-based threats. The maturity model below is my attempt to infuse an AI perspective into existing maturity models.

CONCLUSION

When many enterprises and organizations think about AI and ML, they tend to focus on how this new technology can help them innovate, improve productivity and reduce operational costs. The challenges and potential minefields that emanate from the use of the same technology by cyber threat actors is either not clearly understood or in some cases minimized outright. This research study has served to inform and educate cybersecurity management practitioners and researchers about potential pitfalls and emerging cyber risks and equip them to deal adequately with the emerging AI and ML-driven adversarial threats. The reader needs to carefully evaluate how this research is applicable to their own context and situations.

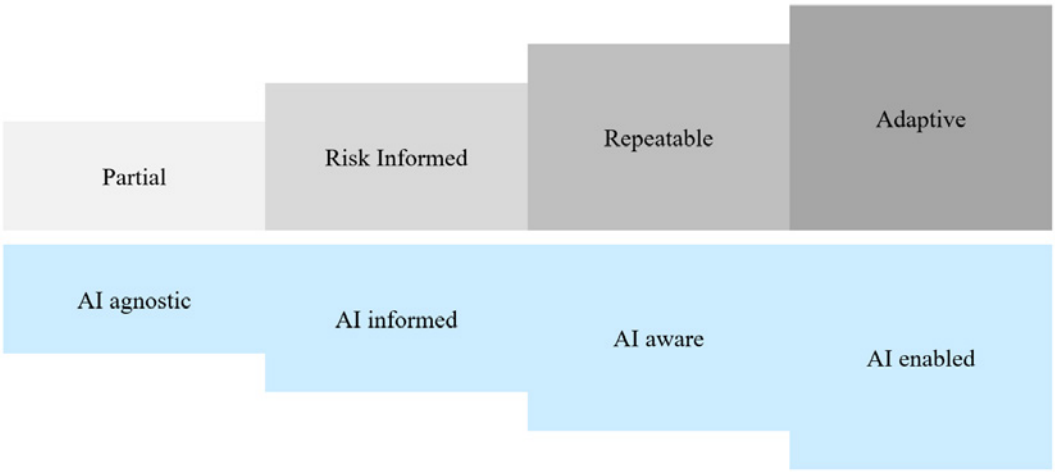

Figure 2 *– NIST CSF Tiers for cybersecurity risk management (adapted for AI-driven adversarial attacks by Author - National Institute of Standards and Technology (2024))*

REFERENCES

Agzayal, Y., & Bouhorma, M. (2024). AI-Driven Cyber Risk Management Framework. In: Ben Ahmed, M., Boudhir, A.A., El Meouche, & R., Karas, İ.R. (eds) Innovations in Smart Cities Applications Volume 7. SCA 2023. Lecture Notes in Networks and Systems, vol 906. Springer, Cham. httpss://doi.org/10.1007/978-3-031-53824-7_51

Asatiani, A., Malo, P., Nagbøl, P. R., Penttinen, E., Rinta-Kahila, T., & Salovaara, A., (2020) "Challenges of Explaining the Behavior of Black-Box AI Systems," MIS Quarterly Executive: Vol. 19 : Iss. 4 , Article 7. httpss://aisel.aisnet.org/misqe/vol19/iss4/7

Aziz, S. & Dowling, M. (2019). Machine Learning and AI for Risk Management. In: Lynn, T., Mooney, J., Rosati, P., Cummins, M. (eds) Disrupting Finance. Palgrave Studies in Digital Business & Enabling Technologies. Palgrave Pivot, Cham. httpss://doi.org/10.1007/978-3-030-02330-0_3

Carabantes, M. (2020) Black-box artificial intelligence: an epistemological and critical analysis. *AI & Soc* **35**, 309–317 (2020). httpss://doi.org/10.1007/s00146-019-00888-w

Chaddad A, Peng J, Xu J, Bouridane A. Survey of Explainable AI Techniques in Healthcare. *Sensors*. 2023; 23(2):634. httpss://doi.org/10.3390/s23020634

Chakravorti, B. (2024), AI's Trust Problem - Twelve persistent risks of AI that are driving skepticism, httpss://hbr.org/2024/05/ais-trust-problem

Foysal, A., Islam, S.M., & Rahaman, T. (2019). Classification of AI Powered Social Bots on Twitter by Sentiment Analysis and Data Mining through SVM. *International Journal of Computer Applications, 177*, 13-19.

Fritsch, L., Jaber, A., & Yazidi, A. (2022). An Overview of Artificial Intelligence Used in Malware. In: Zouganeli, E., Yazidi, A., Mello, G., Lind, P. (eds) Nordic Artificial Intelligence Research and Development. NAIS 2022. Communications in Computer and Information Science, vol 1650. Springer, Cham. httpss://doi.org/10.1007/978-3-031-17030-0_4

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, *36*(1). httpss://doi.org/10.1080/08839514.2022.2037254

Hart, D. (2023). Uncovering How AI's Dual Relationship With Cybersecurity Operates, httpss://www.forbes.com/councils/forbestechcouncil/2023/06/28/uncovering-how-ais-dual-relationship-with-cybersecurity-operates/

Kaloudi, N. and Li, J. 2020. The AI-Based Cyber Threat Landscape: A Survey. ACM Comput. Surv. 53, 1, Article 20 (January 2021), 34 pages. httpss://doi.org/10.1145/3372823

Kumar, P. Adversarial attacks and defenses for large language models (LLMs): methods, frameworks & challenges. *Int J Multimed Info Retr* **13**, 26 (2024). httpss://doi.org/10.1007/s13735-024-00334-8

Ishai R., Asaf S., Yuval E., & Lior R. (2021). Adversarial Machine Learning Attacks and Defense Methods in the Cyber Security Domain. ACM Comput. Surv. 54, 5, Article 108 (June 2022), 36 pages. httpss://doi.org/10.1145/3453158

Lyu, S. (2024). "DeepFake the menace: mitigating the negative impacts of AI-generated content", Organizational Cybersecurity Journal: Practice, Process and People, Vol. ahead-of-print No. ahead-of-print. httpss://

doi.org/10.1108/OCJ-08-2022-0014

Malatji, M., & Tolah, A. (2024). Artificial intelligence (AI) cybersecurity dimensions: a comprehensive framework for understanding adversarial and offensive AI. *AI Ethics*. httpss://doi.org/10.1007/s43681-024-00427-4

Filiz Mizrak (2023). Integrating cybersecurity risk management into strategic management: a comprehensive literature review. Research Journal of Business and Management (RJBM), 10(3), 98-108. https://doi.org/10.17261/Pressacademia.2023.1807

Minh, D., Wang, H.X., Li, Y.F. et al. Explainable artificial intelligence: a comprehensive review. Artif Intell Rev 55, 3503–3568 (2022). httpss://doi.org/10.1007/s10462-021-10088-y

National Institute of Standards and Technology (2024) The NIST Cybersecurity Framework (CSF) 2.0. (National Institute of Standards and Technology, Gaithersburg, MD), NIST Cybersecurity White Paper (CSWP) NIST CSWP 29. httpss://doi.org/10.6028/NIST.CSWP.29

Núñez, F., Esteban, J. (2022). Adversarial machine learning for cyber security https://hdl.handle.net/2117/372347

Radanliev, P., De Roure, D., Page, K., Nurse, J.R.C., Mantilla M. R., Santos, O., Maddox, L., & Burnap, P. (2020). Cyber risk at the edge: current and future trends on cyber risk analytics and artificial intelligence in the industrial internet of things and industry 4.0 supply chains, Cybersecurity, Volume 3, Article number 13 (2020). httpss://doi.org/10.1186/s42400-020-00052-8

Salem, A.H., Azzam, S.M., Emam, O.E., & Abohany, A., A. (2024). Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. *J Big Data* 11, 105. httpss://doi.org/10.1186/s40537-024-00957-y

Shashkov A, Hemberg E, Tulla M, O'Reilly U-M. Adversarial agent-learning for cybersecurity: a comparison of algorithms. *The Knowledge Engineering Review*. 2023;38:e3. doi:10.1017/S0269888923000012

Schreiber, A. & Schreiber, I. (2024). Bridging knowledge gap: the contribution of employees' awareness of AI cyber risks comprehensive program to reducing emerging AI digital threats, *Information and Computer Security*, Vol. ahead-of-print No. ahead-of-print. httpss://doi.org/10.1108/ICS-10-2023-0199

Van Haastrecht, M et. al. 2021. Respite for SMEs: A systematic review of sociotechnical cybersecurity metrics. Applied Sciences (Switzerland) 11, 15 (8 2021), 6909. httpss://doi.org/10.3390/app11156909

Waizel, G. (2024). Bridging the AI divide: The evolving arms race between AI-driven cyber attacks and AI-powered cybersecurity defenses. In *International Conference on Machine Intelligence & Security for Smart Cities (TRUST) Proceedings* (Vol. 1, pp. 141-156). httpss://www.scrd.eu/index.php/trust/article/view/554

Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, M., Elovici, Y., & Biggio, B. (2023). The Threat of Offensive AI to Organizations, Computers & Security, Volume 124, 2023, 103006, ISSN 0167-4048, httpss://doi.org/10.1016/j.cose.2022.103006

Wang, Y *et al*. (2023), "Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey," in *IEEE Communications Surveys & Tutorials*, vol. 25, no. 4, pp. 2245-2298, Fourthquarter 2023, httpss://doi.org/10.1109/COMST.2023.3319492

Yamin, M. M., Ullah, M., Ullah, H., & Katt,B. (2021). Weaponized AI for cyber attacks,

Yang, W., Wei, Y., Wei, H. *et al.* Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Hum-Cent Intell Syst* **3**, 161–188 (2023). httpss://doi.org/10.1007/s44230-023-00038-y