How AI Shapes Cyber Risk Management

Sander Zeijlemaker
Cyber Security at MIT Sloan, Sloan School of Management, Massachusetts Institute of Technology, Boston, USA
szeijl@mit.edu

Yaphet Lemiesa
Cyber Security at MIT Sloan, Sloan School of Management, Massachusetts Institute of Technology, Boston, USA

Saskia Laura Schröer
University of Liechtenstein

Michael Siegel
Cyber Security at MIT Sloan, Sloan School of Management, Massachusetts Institute of Technology, Boston, USA

Keywords: Cybersecurity, Artificial Intelligence, systemic structure change, qualitative modeling

Extended Abstract: Technological evolution has ushered in an era of profound digital transformation. As society increasingly embraces smart cities, Industry 4.0, and e-health infrastructures, its exposure to cyber threats escalates. Although cybersecurity technologies have advanced, adversaries now wield AI to scale attacks with unprecedented speed, stealth, and sophistication (FBI, 2024; Aggrey et al., 2024; Schröer et al., 2025). In 2024, cybercrime became the world's third-largest economy, valued at $9.5 trillion (Bloomberg, 2024). This work explores the changing cyber risk landscape by identifying the systemic structures that underlie cyber risk management and illustrating how AI disrupts, strengthens, or redefines these structures.

Methodology and Framework

This analysis synthesizes literature research, expert interviews, and three executive workshops. Central to the study is a dynamic, systemic modeling approach supported by a walkthrough of the Colonial Pipeline incident via the cyber-kill-chain framework (Zeng & Germanos, 2019). Systemic archetypes such as the "aging chain," "capability trap" (Repenning & Sterman, 2002), and "success-to-success feedback" underpin the analysis, offering causal loop insights into how prevention, detection, response, and recovery efforts evolve amid AI-driven change.

Findings: Systemic Structures of Cyber Risk Management

Cyber systems transition through three primary states—secure, at risk, and compromised (Zeijlemaker & Siegel, 2023; Estay, 2021; Armenia et al., 2021; Jalali et al., 2019). The systemic model reveals that prevention, detection, response, and recovery form core defensive capabilities that influence these transitions through interlinked balancing loops (CIS, 2021; GCIO, 2020; Muneer, 2021; Pascoe 2023; Zeijlemaker & Siegel, 2023; Armenia et al., 2021; Jalali et al., 2019). These efforts mitigate adversarial threats but also introduce cost pressures, leading to reinforcing loops that reflect trade-offs between security investments and organizational profitability (Eling et al., 2021; Paté-Cornell et al., 2018; Zeijlemaker et al., 2023; Zeijlemaker & Siegel, 2023; Armenia et al., 2021; Jalali et al., 2019). When left unaddressed, system vulnerabilities compound over time, especially in resource-constrained environments. The lateral movement of ransomware and other advanced threats forms an accelerating feedback loop, as compromised systems infect adjacent assets, amplifying organizational exposure (Meurs et al., 2023; Luo & Liao, 2009).

This dynamic is further complicated by the "capability trap," in which short-term operational pressures defer investment in essential security functions such as patching, monitoring, and staff training (Zeijlemaker & Siegel, 2023). Over time, this undermines the maturity and effectiveness of cyber risk management programs. Survey data show that 53% of incidents stem from security control failures, while 47% result from unintended control lapses (IBM Security, 2022), confirming the systemic fragility induced by reactive governance models.

Ransom payments introduce additional risk loops, such as repeat targeting and incomplete data recovery (Cybereason, 2022).

The Colonial Pipeline Case: A Systemic Perspective

The Colonial Pipeline incident exemplifies how attackers traverse the kill chain via successive, systemically linked phases—from reconnaissance and weaponization to exfiltration and ransom deployment. The attack exploited a single reused credential via a VPN lacking multi-factor authentication followed by lateral movement, domain control compromise, and critical data exfiltration. The breach inflicted over $5 billion in economic losses and exposed gaps in detection, segmentation, and contingency planning (Beerman et al., 2023; Brash, 2021). In the era of AI, each of these phases can be enhanced by AI in terms of speed, volume, and advanced automation from both adversarial and defenders' perspectives (Zhang et al., 2024).

AI's Dual Role: Disruption and Defense

AI intensifies cyber risks by enabling autonomous attack agents (The Grungq, 2017), automated exploit generation, polymorphic malware (e.g., Black Mamba-Montalbano, 2023), and federated botnets (Ilascu, 2024). Conversely, AI enhances defense through the following mechanisms:

- Automated Security Hygiene: This includes self-patching systems (Sibanda, 2024), self-healing software code (ABN, 2021), self-driving trustful networks (Hireche et al., 2022), adaptive and autonomous identity and access management (The Hacker News, 2024), and continuous attack surface management systems (Vindhya et al., 2023) with automated threat mitigation (Komaragiri & Edward, 2023).

- Deceptive Defense Systems: These leverage moving target defense that dynamically alters system configurations (Jajodia et al., 2019) and complicates attacks while affecting adversaries' cognitive biases (Bilinski et al., 2020).

- Autonomous Detection and Response: Enabled by SOAR (Cyware, 2023) and XDR platforms (Joshi, 2024), GAN-enhanced simulation (Shehu et al., 2023), and AI-guided incident forensics (Musman et al., 2019).

Three AI-induced systemic structure changes emerged:

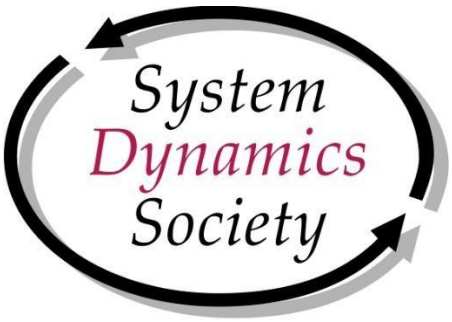
1. Deception Capabilities: Defenders can now create controlled compromised systems that act as honeypots or decoys, misleading adversaries while avoiding real-world impact. This activity introduces new balancing and reinforcing loops involving deception capabilities and adversary recalibration (Jajodia et al., 2019).

2. Two-Step Success Feedback: Attackers first poison AI-based security systems (e.g., whitelisting or model evasion) and then exploit infrastructure vulnerabilities. These coordinated strategies create cascading success loops, requiring defenders to retrain models and harden detection protocols continuously.

3. Self-Replicating AI Exploits: Misused generative AI tools (e.g., LLMs) can autonomously spawn additional attack vectors as long as they remain undetected, further reinforcing this dynamic. This method represents a major paradigm shift in which AI becomes both an attacker and a target surface (Cohen et al., 2024).

Conclusion

Our findings illustrate that AI is not merely enhancing traditional cyberattack methods; it is redefining the underlying systemic architecture of cyber risk management. Through a fusion of historical incident analysis, expert consultation, and literature synthesis, we confirm that while AI exacerbates existing tensions between security investment and risk exposure, it also presents novel structural opportunities. The discovery of three new systemic feedback structures demands urgent attention from strategists, CISOs, and policymakers. Failure to account for these shifts may leave organizations vulnerable to increasingly autonomous, deceptive, and persistent adversaries. A proactive, adaptive, and systemic approach to AI-powered cybersecurity is essential.



References

ABN. (2021, March 19). ABN AMRO first buyer of innovative self-healing cybersecurity software. ABN AMRO. https://www.abnamro.com/en/news/abn-amro-first-buyer-of-innovative-self-healing-cybersecurity-software

Aggrey, R., Adjei, B. A., Afoduo, K. O., Dsane, N. A. K., Anim, L., & Ababio, M. A. (2024). Understanding and mitigating AI-powered cyber-attacks. International Journal for Multidisciplinary Research (IJFMR), 6(6). https://doi.org/10.36948/ijfmr.2024.v06i06.33563

Armenia, S., Angelini, M., Nonino, F., Palombi, G., & Schlitzer, M. F. (2021). A dynamic simulation approach to support the evaluation of cyber risks and security investments in SMEs. Decision Support Systems, 147, 113580. https://doi.org/10.1016/j.dss.2021.113580

Beerman, J., Berent, D., Falter, Z., & Bhunia, S. (2023, May). A review of Colonial Pipeline ransomware attack. In 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW). https://sbhunia.me/publications/manuscripts/ccgrid23.pdf

Bilinski, M., diVita, J., Ferguson-Walter, K., Fugate, S., Gabrys, R., Mauger, J., & Souza, B. (2020, January 1). Lie another day: Demonstrating bias in a multi-round cyber deception game of questionable veracity. Springer International Publishing. https://link.springer.com/chapter/10.1007/978-3-030-64793-3_5

Bloomberg. (2024, April 22). The world's third-largest economy has bad intentions — and it's only getting bigger. Bloomberg. https://sponsored.bloomberg.com/quicksight/check-point/the-worlds-third-largest-economy-has-bad-intentions-and-its-only-getting-bigger

Brash, R. (2021, May 11). Lessons learned from the Colonial Pipeline attack. Control Engineering. https://www.industrialcybersecuritypulse.com/facilities/lessons-learned-from-the-colonial-pipeline-attack/
Centre of Internet Security (CIS). (2021). CIS controls V8. East Greenbush, NY: CIS.

Cohen, S., Bitton, R., & Nassi, B. (2024). Here comes the AI worm: Unleashing zero-click worms that target GenAI-powered applications. arXiv Preprint arXiv:2403.02817.

Cyware. (2023, December 1). SOAR and AI in cybersecurity: Reshaping your security operations. Cyware. https://cyware.com/security-guides/security-orchestration-automation-and-response/from-insight-to-action-how-ai-and-soar-are-reshaping-security-operations-13d9

Cybereason. (2022). Ransomware: The True Cost to Business 2022. Report. https://www.cybereason.com/ransomware-the-true-cost-to-business-2022.

Eling, M., McShane, M., & Nguyen, T. (2021). Cyber risk management: History and future research directions. Risk Management & Insurance Review, 24, 93–125. https://doi.org/10.1111/rmir.12169

Estay, D. (2021). A system dynamics, epidemiological approach for high-level cyber-resilience to zero-day vulnerabilities. Journal of Simulation, 17(1-16). https://doi.org/10.1080/17477778.2021.1890533

Federal Bureau of Investigation (FBI). (2024, May 8). FBI warns of increasing threat of cyber criminals utilizing artificial intelligence. FBI San Francisco. https://www.fbi.gov/contact-us/field-offices/sanfrancisco/news/fbi-warns-of-increasing-threat-of-cyber-criminals-utilizing-artificial-intelligence

Government Chief Information Officer (GCIO). (2020). An overview of ISO/IEC 27000 family of Information Security Management System Standards. Office of the Government Chief Information Officer. (Original work published April 2015, updated May 2020).

Hireche, O., Benzaïd, C., & Taleb, T. (2022). Deep data plane program
mming and AI for zero-trust self-driven networking in beyond 5G. Computer Networks, 203, 108668. https://doi.org/10.1016/j.comnet.2021.108668

IBM Security. (2022). Cost of a data breach report 2022. https://www.ibm.com/reports/data-breach

Ilascu, I. (2024, April 10). Malicious PowerShell script pushing malware looks AI-written. Bleeding Computer. https://www.bleepingcomputer.com/news/security/malicious-powershell-script-pushing-malware-looks-ai-written/

Jalali, M. S., Siegel, M., & Madnick, S. (2019). Decision-making and biases in cybersecurity capability development: Evidence from a simulation game experiment. The Journal of Strategic Information Systems, 28(1), 66–82. https://doi.org/10.1016/j.jsis.2018.09.003

Jajodia, S., Cybenko, G., Liu, P., Wang, C., & Wellman, M. (Eds.). (2019). Adversarial and uncertain reasoning for adaptive cyber defense: Control-and game-theoretic approaches to cybersecurity (Vol. 11830). Springer Nature.

Joshi, J. (2024, May 13). The impact of AI on endpoint detection and response. Proficio. https://www.proficio.com/blog/ai-endpoint-detection-and-response-edr/

Komaragiri, V. B., & Edward, A. (2022). AI-driven vulnerability management and automated threat mitigation. International Journal of Scientific Research and Management (IJSRM). https://www.semanticscholar.org/reader/76e0d0a41ff2d7e7b8f38b52ce10c6a3eea5c613

Luo, X., & Liao, Q. (2009). Ransomware: A new cyber hijacking threat to enterprises. In Handbook of Research on Information Security and Assurance (pp. 1–6). IGI Global. https://doi.org/10.4018/978-1-59904-855-0.ch001

Meurs, T., Cartwright, E., Cartwright, A., Junger, M., Hoheisel, R., Tews, E., & Abhishta, A. (2023, November). Ransomware economics: A two-step approach to model ransom paid. In 2023 APWG Symposium on Electronic Crime Research (eCrime) (pp. 1–13). IEEE.

Montalbano, E. (2023, March 8). AI-powered 'BlackMamba' keylogging attack evades modern EDR security. DarkReading. https://www.darkreading.com/endpoint-security/ai-blackmamba-keylogging-edr-security

Muneer, F. (2021). Cybersecurity capability maturity model, version 2.0. U.S. Department of Energy.
Musman, S., Booker, L., Applebaum, A., & Edmonds, B. (2019, May 10). Steps toward a principled approach to automating cyber responses. SPIE Digital Library. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/11006/2518976/Steps-toward-a-principled-approach-to-automating-cyber-responses/10.1117/12.2518976.full#=

Pascoe, C. E. (2023). Public draft: The NIST Cybersecurity Framework 2.0.

Paté-Cornell, M. E., Kuypers, M., Smith, M., & Keller, P. (2018). Cyber risk management for critical infrastructure: A risk analysis model and three case studies. Risk Analysis, 38(2), 226–241.

Repenning, N. P., & Sterman, J. D. (2002). Capability traps and self-confirming attribution errors in the dynamics of process improvement. Administrative Science Quarterly, 47, 265–295.

Schröer, S.L., Apruzzese, G., Human, S., Laskov, P., Anderson, H. S, Bernroider, E. W. N., Fass, A., Nassi, B., Rimmer, V., Roli, F., Salam, S., Shen, A, Sunyaev, A, Wadhwa-Brown, T., Wagner, I., Wang, G. (2025) SoK: On the Offensive Potential of AI. IEEE Conference on Secure and Trustworthy Machine Learning.

Shehu, A., Umar, M., & Aliyu, A. (2023). Cyber kill chain analysis using artificial intelligence. Asian Journal of Research in Computer Science, 16(3), 210–219. https://doi.org/10.9734/ajrcos/2023/v16i3357

Sibanda, I. (2024, November 29). Automated patch management: A proactive way to stay ahead of threats. ComputerWeekly. https://www.computerweekly.com/feature/Automated-patch-management-A-proactive-way-to-stay-ahead-of-threats

The Grungq. (2017). The Triple A Threat: Aggressive autonomous agents. BlackHat. https://www.blackhat.com/docs/webcast/12142017-the-triple-a-threat.pdf

The Hacker News. (2024, November 15). How AI is transforming IAM and identity security. The Hacker News. https://thehackernews.com/2024/11/how-ai-is-transforming-iam-and-identity.html

Vindhya, L., Mahima, B. Gowda, Gowramma Gaari Sindhu, & Keerthan, V. (2023). International Journal of Advanced Research in Science, Communication and Technology (IJARSCT, 3(8), April 2023. https://ijarsct.co.in/Paper9533.pdf

Zeijlemaker, S., & Siegel, M. (2023, January 3–6). Capturing the dynamic nature of cyber risk: Evidence from an explorative case study [Conference session]. Hawaii International Conference on System Sciences (HICSS)–56, Hawaii.

Zeijlemaker, S., Hetner, C., & Siegel, M. (2023, June). Four areas of cyber risk that boards need to address. Harvard Business Review. https://hbr.org/2023/06/4-areas-of-cyber-risk-that-boards-need-to-address

Zeng, W., & Germanos, V. (2019). Modelling hybrid cyber kill chain. PNSE@Petri Nets/ACSD. https://www.semanticscholar.org/paper/Modelling-Hybrid-Cyber-Kill-Chain-Zeng-Germanos/f5cb1f80c669562d3dd61b4dcbc6410a5d015c62

Zhang, C., Pal, R., Nicholson, C., & Siegel, M. (2024, December). (Gen) AI Versus (Gen) AI in Industrial Control Cybersecurity. In 2024 Winter Simulation Conference (WSC) (pp. 2739-2750). IEEE.