Assignment 6.1: Automating AI Security Risks

Joshua A. Wortz

University of San Diego

CSOL-530

Cameron Carter, MS

06/015/2025

Artificial intelligence (AI) is transforming the way organizations approach cybersecurity. With its ability to sift through massive amounts of data and identify unusual behavior in real time, AI can help security teams stay ahead of threats that would otherwise go unnoticed. By automating repetitive tasks and identifying risks more quickly than humans alone can, AI enhances both the speed and accuracy of security operations. Still, using AI for cybersecurity is not without complications. There are ethical concerns regarding privacy, bias, and transparency, as well as operational risks such as false alerts or overreliance. This paper examines the benefits and challenges of utilizing AI in security, as well as strategies organizations can employ to mitigate potential downsides.

One of the primary strengths of AI in cybersecurity is its ability to automate tasks. Security tools that use machine learning can analyze logs and network activity much faster than human analysts. Platforms like Splunk, when enhanced with algorithms such as Random Forest or XGBoost, can detect complex threats, including advanced persistent threats (APTs), with high accuracy (Ali et al., 2025). These tools do not just flag issues—they can sort alerts by severity, initiate responses, and even isolate affected systems without needing immediate human input. Unsupervised learning models are also being used to detect unusual user behavior, which is key for spotting insider threats or compromised accounts.

Another significant benefit is real-time analysis. Traditional monitoring tools often struggle to keep up with the volume of data generated by modern IT environments. AI-enabled systems can scan thousands of events per second, allowing security teams to respond almost instantly. This speed is crucial for preventing data breaches and limiting the damage from ransomware and other rapidly evolving attacks. Plus, by automating basic tasks, AI helps reduce burnout among security analysts, who would otherwise be overwhelmed by the volume of alerts.

Despite these advantages, the use of AI for cybersecurity also raises significant ethical concerns. For example, AI systems often require access to detailed logs and user data to operate effectively. This can put sensitive information at risk and create concerns about digital surveillance and consent (Taddeo & Floridi, 2018). Additionally, AI models can reflect biases found in their training data. If the data is skewed—say, by overrepresenting certain types of users or regions—the system might unfairly target or ignore specific behaviors (Veale & Binns, 2017).

Another problem is transparency. Many AI systems are challenging to interpret and are profound learning models. When a system flags a threat or makes a decision, it is not always clear why. That lack of insight can make it harder to investigate incidents or explain actions during audits.

There are also risks associated with relying too heavily on AI for cybersecurity. One common issue is the rate of false positives—alerts triggered by harmless activity. If this happens too frequently, analysts may begin to ignore alerts altogether. On the flip side, false negatives are just as dangerous, since they let real threats slip through. Overreliance on automation can also diminish the role of human expertise, which remains crucial when responding to novel or complex attacks. Moreover, like any software, AI systems can be exploited. Attackers may attempt to feed misleading data to deceive the system or launch adversarial attacks designed to evade detection (Brundage et al., 2018).

To make AI safer and more reliable, organizations should take steps to limit these risks. One approach is to keep a human in the loop for critical decisions. Allowing AI to handle initial triage is acceptable, but final actions, such as blocking traffic or removing files, should require human approval. Regular testing is also important. AI models must be regularly retrained and

continually evaluated against emerging threats to remain effective. Some teams use adversarial testing or red-teaming to simulate attacks and spot weaknesses (Sommer & Paxson, 2010).

Another helpful practice is using explainable AI (XAI) tools. These tools give insight into how the system made a decision, which is helpful for audits and helps build trust. Organizations should also create internal policies that address data privacy, bias checks, and access controls. This type of governance ensures that AI is used responsibly and in compliance with legal requirements, such as the GDPR or recent federal AI guidance (White House, 2023). Together, these steps help strike a balance between innovation and caution.

Artificial intelligence (AI) is quickly becoming a powerful component of modern cybersecurity. Its ability to detect threats, automate responses, and manage vast volumes of data makes it a valuable asset in defending against increasingly complex attacks. However, these benefits come with real concerns that need to be addressed. By being thoughtful about how AI is used, especially in terms of ethics, reliability, and transparency, organizations can maximize the benefits of these tools while minimizing the risks. The key is finding the right balance between automation and human judgment.

References

Ali, G., Shah, S., & ElAffendi, M. (2025). Enhancing cybersecurity incident response: AI-driven optimization for strengthened APT detection. *Results in Engineering, 25*, 104078. https://doi.org/10.1016/j.rineng.2025.104078

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., ... & Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv. https://arxiv.org/abs/1802.07228

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

Taddeo, M., & Floridi, L. (2018). How AI can be a force for good. *Science, 361*(6404), 751-752. https://doi.org/10.1126/science.aat5991

Veale, M., & Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society, 4*(2). https://doi.org/10.1177/2053951717743530

White House. (2023). *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-safe-secure-and-trustworthy-ai/