The Dangers of Integrating Autonomous Artificial Intelligence into US Department of Defense Systems

Jessica D. Wrzesien

College of Computing, Georgia Institute of Technology

CS 3001 Computing & Society

Dr. Clint Zeagler

April 25, 2024

Introduction

In this paper, I will explore and answer the question, "How could autonomous artificial intelligence be integrated into the United States Department of Defense and what dangers could arise from doing so?" As artificial intelligence, henceforth referred to as AI, gains further exposure, the risks associated with its utilization in critical systems become increasingly evident. Despite these risks, the government *must* begin using AI in their defense systems due to the growing threat of AI-powered malware (Schram, 2021). So the question stands, are the dangers of AI use in defense systems greater than the risks posed by AI-based cyberattacks? In this paper, I will argue that they are not, as AI-driven defense systems are crucial for combating AI-based attacks (Schram, 2021). Hence, it is important to understand the risks created by the use of this essential AI so as to develop protective measures that allow for its most efficient use.

The use of AI in cyber operations is a well established concept. Because of its many uses in the ever evolving cyber defense field, it has been used and developed by private companies who license their software for many years. It is effective in completing repetitive tasks, monitoring network behavior, detecting malicious activity, improving endpoints, etc. (Schram, 2021). However, this paper shall focus specifically on autonomous AI due to the risky but immensely useful applications of an AI that is able to adapt and respond to novel situations and contexts with little to no human involvement (Danks & London, 2017).

United States Department of Defense Intentions

The United States Department of Defense, henceforth referred to as the DoD, began publicly expressing its intentions to further research and eventually implement AI into their systems back in 2018 with the release of the "2018 Artificial Intelligence Strategy", which aimed

to “(1) assess the state of AI relevant to DoD and address misconceptions; (2) carry out an independent introspective assessment of the posture of DoD in AI; and (3) develop recommendations for internal actions, external engagements, and legislative actions to improve DoD’s posture in AI,” (Tarraf et al., 2019). In April 2023, the DoD released the memorandum of their “2022 Software Modernization Strategy”, confirming that the government is certainly aiming to improve the software in their defense systems. It states that the government plans to adopt cloud systems and make use of more enterprise services, which will likely include the aforementioned AI-based cyber defense software from the private sector (US DoD, 2022). Then, the DoD announced its intentions to accelerate the adoption of AI with the “2023 Data, Analytics and Artificial Intelligence Adoption Strategy” (Clark, 2024). This strategy outlines the DoD’s initiatives to promote speed, delivery, learning, and responsible development of AI across the DoD. It also acknowledges the DoD’s inability to make use of more enterprise services as stated in their Modernization Strategy as they do not align with the department’s ethical guidelines (US DoD, 2023). Thus, it is clear that the DoD is actively striving to improve their cyber defense and other military systems to be capable of protecting against advanced AI-driven malware.

Autonomous AI in Threat Detection and Analysis

In the context of the cyber operations of the DoD, threat detection and analysis can be defined as the process of identifying, assessing, and evaluating potential threats, vulnerabilities, and weaknesses within the DoD's information systems, networks, and infrastructure (Tarraf et al., 2019).

Applications. Autonomous AI systems can be trained to continuously monitor network traffic and system logs for anomalous behavior or indicators of malicious activity more

effectively than traditional methods (Bozic, 2023b; Tarraf et al., 2019). These systems can use machine learning algorithms to analyze vast amounts of data in real-time, identify patterns indicative of potential threats, and generate alerts or responses without human intervention. For example, autonomous AI algorithms can detect abnormal patterns of network traffic that may indicate a cyberattack, such as a distributed denial-of-service (DDoS) attack or data exfiltration attempt (Danks & London, 2017).

Additionally, autonomous AI can be utilized in intrusion detection systems (IDS) and intrusion prevention systems (IPS) to identify and block unauthorized access attempts or malicious software by enabling adaptive and proactive defense mechanisms (Bozic, 2023b; Tarraf et al., 2019). AI-driven IDS/IPS systems can learn from historical attack data to improve their detection accuracy and identify previously unknown threats. Moreover, autonomous AI can enable dynamic adjustments to security policies and configurations in response to evolving threats, minimizing the risk of false positives and false negatives (Scrham, 2021).

Autonomous AI can also be used in the conduction of vulnerability assessments and penetration testing such as fuzz testing by automating the discovery, analysis, and remediation of security weaknesses in DoD systems and infrastructure that could be exploited by adversaries (Bozic, 2023b; Tarraf et al., 2019). AI-based vulnerability scanners can autonomously identify and prioritize vulnerabilities based on their severity and potential impact on critical assets. Furthermore, autonomous AI agents can simulate sophisticated cyberattacks and penetration testing scenarios to assess the resilience of DoD networks and systems against various threats. By leveraging AI, the DoD can conduct more comprehensive and continuous security assessments, thereby reducing the window of exposure to potential cyber threats (Schram, 2021).

As mentioned in the "2022 Software Modernization Strategy", the DoD plans to employ cloud systems as a part of their development efforts. Hence, autonomous AI could be utilized in cloud security for threat analysis by continuously monitoring cloud environments, analyzing vast amounts of data to identify suspicious activities or anomalies, and autonomously responding to potential security threats in real-time (Bozic, 2023b; Tarraf et al., 2019). This includes detecting unauthorized access attempts, abnormal user behaviors, or potential data breaches, and taking proactive measures to mitigate threats. By leveraging autonomous AI in cloud security, the DoD can enhance its ability to detect and respond to cybersecurity threats effectively, thereby improving the overall resilience of its cloud infrastructure.

Risks. There are several challenges of autonomous AI that exist no matter its application. They include unsupervised learning, algorithmic biases such as political and demographic bias, inscrutability, and over-dependence (Bozic, 2023a; "Dangers in the Machine", n.d.; Danks & London, 2017; Schram, 2021). While these are not the only problems involved when implementing autonomous AI, for the sake of brevity this paper will focus only on these four.

Defensive autonomous AI systems often use unsupervised learning, which occurs when an algorithm learns from unlabeled data without human supervision, in order to completely understand its network environment. While unsupervised learning can uncover previously unknown threats, it may also produce false positives or false negatives if the algorithm misinterprets benign activities as malicious or overlooks subtle indicators of real threats. This can result in wasted resources and unnecessary disruptions to operations or, conversely, in critical threats being overlooked (Schram, 2021). Additionally, unsupervised learning AI models could fall prey to data poisoning attacks in which an attacker could inject anomalous data points

into an AI's training data to bias the representation learned by an autoencoder, for example (Shen & Xia, 2020).

Although not as common in unsupervised AI systems, in supervised AI systems it is possible for autonomous AI systems to inherit and perpetuate algorithmic biases, leading to discriminatory outcomes in threat detection (Bozic, 2023a; Danks & London, 2017). For example, if the training data used to develop AI algorithms contains biases related to political affiliations or demographic characteristics, the AI may exhibit biased behavior in identifying threats, potentially overlooking or overemphasizing certain types of cyber threats based on biased patterns in the data. This could result in unequal protection for different groups or regions, undermining the fairness and effectiveness of threat detection efforts ("Danger in the Machine", n.d).

Over-reliance on autonomous AI for threat detection can lead to complacency and a diminished role for human oversight and intervention. If cybersecurity engineers become overly dependent on AI systems to detect and mitigate threats, they may become less vigilant in monitoring and interpreting security alerts, potentially missing important indicators of emerging cyberattacks or vulnerabilities. Moreover, if AI systems fail to detect sophisticated or novel threats, over-dependence on AI could leave DoD networks and systems vulnerable to exploitation (Bozic, 2023a; Schram, 2021).

Additionally, autonomous AI systems, particularly those using complex machine learning algorithms, may lack transparency and explainability in their decision-making processes. This inscrutability can make it difficult for cybersecurity analysts and engineers to understand how and why certain threat detection decisions are made, especially if they are becoming over reliant on the AI. Without clear explanations for AI-driver threat detection outcomes, it may be

challenging to validate the accuracy and reliability of the results, leading to mistrust in AI systems and potentially overlooking critical threats or vulnerabilities (Bozic, 2023a; Schram, 2021).

Autonomous AI in Threat Mitigation

In the context of the cyber operations of the DoD, threat mitigation can be defined as the implementing measures to reduce or eliminate threats and vulnerabilities to DoD information networks, systems, and infrastructure (Schram, 2021).

Applications. As previously stated, autonomous AI is able to detect and analyze malicious actors by continuously monitoring network traffic and system logs, powering intrusion detection and prevention systems (IDPS), conducting vulnerability assessments and penetration testing, and analyzing data from cloud systems. Similarly, autonomous AI can also be used to resolve any threats or abnormalities discovered during those processes. In IDPSs, Autonomous AI can be used to autonomously respond to detected intrusions by initiating real-time countermeasures such as blocking suspicious IP addresses, isolating compromised systems, or adjusting firewall rules to prevent further unauthorized access. In vulnerability management, it can be used to automatically prioritize and remediate vulnerabilities discovered during vulnerability assessments and penetration testing by applying patches, updates, or configuration changes to vulnerable systems and applications. Lastly, Autonomous AI can enhance cloud security by autonomously detecting and responding to threats or abnormalities in cloud systems, such as unauthorized access attempts, data breaches, or configuration errors (Bozic, 2023b; “Risk of AI”, n.d.). This may include automatically isolating compromised cloud instances, encrypting sensitive data, or adjusting access controls to mitigate threats in cloud environments.

Risks. Just as in its use in threat detection and analysis, autonomous AI introduces the issues of unsupervised learning, algorithmic biases such as political and demographic bias, inscrutability, and over-dependence when used in threat mitigation.

While it is true that using autonomous AI could help in minimizing the detection of false positives and false negatives (Schram, 2021), the opposite is also true in that AI systems may incorrectly classify benign activities as threats (false positives) or fail to detect genuine threats (false negatives) that humans would not. This can lead to unnecessary blocking of legitimate users or activities, disrupting operations, or allowing malicious actors to evade detection and continue their activities, undermining the effectiveness of threat mitigation efforts.

Furthermore, autonomous actions taken by AI systems without human oversight may have unintended consequences or side effects. For example, blocking suspicious IP addresses could inadvertently disrupt legitimate network traffic, causing service disruptions or impacting mission-critical operations (Schram, 2021). Similarly, applying patches or updates to vulnerable systems without proper testing or validation could introduce new vulnerabilities or system instabilities, exacerbating cybersecurity threats rather than mitigating them. Thus, preventing overreliance in autonomous AI systems by ensuring human surveillance is vital (Bozic, 2023a; Schram, 2021).

However, even AI use *with* human oversight has its disadvantages. As previously stated, autonomous actions taken by AI systems may lack accountability and transparency. Therefore, even if there are personnel overseeing the AI systems, it may be difficult to assess the rationale behind their decisions or hold them accountable for their actions (Bozic, 2023a; Danks & London, 2017). This can undermine trust in AI-driven threat mitigation efforts and hinder efforts to validate the effectiveness and reliability of autonomous responses to cybersecurity threats.

Autonomous actions taken by AI systems in response to cybersecurity threats may also raise legal and compliance concerns, particularly regarding privacy, data protection, and regulatory requirements (Bozic, 2023a). For example, automatically adjusting access controls or encrypting sensitive data without proper authorization or consent could violate privacy laws or contractual obligations, exposing the DoD to legal liabilities or regulatory sanctions (Tarrag et al., 2019). However, as stated in the "2023 Data, Analytics, and Artificial Intelligence Adoption Strategy", one of the reasons the DoD has yet to implement AI is because of such ethical and legal cases, so it is likely that this risk will be highly prioritized to prevent (US DoD, 2023).

Conclusion

The integration of autonomous artificial intelligence (AI) into the United States Department of Defense (DoD) presents both promising opportunities and significant challenges in the field of cyber defense. As the threat landscape evolves and the ability of AI-powered cyberattacks increases, the DoD recognizes the necessity of leveraging AI technologies to enhance its cyber defense capabilities. However, this adoption of autonomous AI in defense systems is not without its dangers. It is essential to recognize that AI is not omnipotent. Despite the benefits of autonomous AI in improving the efficiency and efficacy of cyber defense operations, several risks and challenges must be addressed. These include issues related to unsupervised learning, algorithmic biases, over-dependence, and lack of transparency and accountability in AI-driven decision-making processes.

The risks of giving software unsupervised autonomy must be properly managed and hence understood. Thus, by determining the risks of each application of autonomous AI in

defense cyber operations, solutions in their management and mitigation can be established, allowing for the secure advancement of US cyber defenses.

Recommendations for Future Study

While significant research has been devoted to exploring the utility of autonomous AI within government systems, it is crucial to acknowledge the potential of alternative AI systems. One such approach is discriminative AI, which excels in making precise predictions or classifications based on the characteristics of the input data. By leveraging supervised learning techniques, discriminative AI offers a promising solution to address the limitations associated with unsupervised learning in autonomous AI systems. Moreover, its capability in pattern recognition would be effective in identifying and intercepting malicious actors.

In addition to discriminative AI, further research into the applications of generative AI merits attention. Generative AI, notable for its ability to generate original data based on training samples (Ali, 2024), could be used to facilitate programming and debugging processes for government personnel and contractors, potentially streamlining development efforts and enhancing overall efficiency.

While autonomous AI remains principal in advancing government systems' capabilities, diversifying research efforts to encompass alternative branches of AI such as discriminative and generative AI would unlock new opportunities and fortify cybersecurity measures. Embracing a holistic approach to AI integration will ensure that the DoD, and other government entities, are equipped with the versatile and secure means necessary to combat the evolving challenges in the cyber world.

Bibliography

Ali, S., Ravi, P., Williams, R., DiPaola, D., & Breazeal, C. (2024). Constructing Dreams Using Generative AI. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(21), 23268-23275. https://doi.org/10.1609/aaai.v38i21.30374

Bozic, V. (2023a). The Dangers of Artificial Intelligence. *Research Gate*. https://www.researchgate.net/profile/Velibor-Bozic-2/publication/370659879_THE_DANGERS_OF_ARTIFICIAL_INTELLIGENCE/links/645cc8def43b8a29ba44c49f/THE-DANGERS-OF-ARTIFICIAL-INTELLIGENCE.pdf

Bozic, V. (2023b). AI-Powered Cybersecurity Solutions. *Research Gate*. https://doi.org/10.13140/RG.2.2.35433.47205

Clark, J. (2024). DOD Increases AI Capacity Through Strategy, Alignment. DOD News. *DoD*. https://www.defense.gov/News/News-Stories/Article/Article/3639685/dod-increases-ai-capacity-through-strategy-alignment/#:~:text=The%202018%20DOD%20AI%20Strategy,to%20fielding%20AI%2Denabled%20capabilities.&text=Horowitz%20also%20cited%20DOD%20investments,up%20experimentation%20within%20the%20department

Danger in the Machine: The Perils of Political and Demographic Biases Embedded in AI Systems. (n.d.). *Manhattan Institute*. https://manhattan.institute/article/danger-in-the-machine-the-perils-of-political-and-demographic-biases-embedded-in-ai-systems

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2017/654

Risks of AI & Cybersecurity | Risks of Artificial Intelligence. (n.d.). Malwarebytes. https://www.malwarebytes.com/cybersecurity/basics/risks-of-ai-in-cyber-security#:~:text=Cybersecuri

ty%20professionals%20can%20go%20from

Schram, G. (2021). The Role of Artificial Intelligence in Cyber Operations: An Analysis of AI and Its Application to Malware-Based Cyberattacks and Proactive Cybersecurity (Order No. 28544273). Available from ProQuest Dissertations & Theses A&I; *ProQuest Dissertations & Theses Global*. (2555360198). https://www.proquest.com/dissertations-theses/role-artificial-intelligence-cyber-operations/docview/2555360198/se-2

Shen, J., & Xia, M. (2020). AI Data poisoning attack: Manipulating game AI of Go. ArXiv (Cornell University). https://doi.org/10.48550/arxiv.2007.11820

Tarraf, D. C., Shelton, W. L., Parker, E., Alkire, B., Carew, D., Grana, J., Levedahl, A., Leveille, J., Mondschein, J., Ryseff, J., Wyne, A., Elinoff, D., Geist, E., Harris, B. N., Hui, E., Kenney, C., Newberry, S. J., Sachs, C., Schirmer, P., & Schlang, D. (2019). The Department of Defense posture for artificial intelligence : assessment and recommendations. Rand Corporation.

United States Department of Defense. (2022, February 1). Department of Defense Software Modernization Strategy [PDF]. *DoD*. https://media.defense.gov/2022/Feb/03/2002932833/-1/-1/1/DEPARTMENT-OF-DEFENSE-SOFTWARE-MODERNIZATION-STRATEGY.PDF

United States Department of Defense. (2023). 2023 Data, Analytics, and Artificial Intelligence Adoption Strategy. *DoD*. https://media.defense.gov/2023/Nov/02/2003333300/-1/-1/1/DOD_DATA_ANALYTICS_AI_ADOPTION_STRATEGY.PDF