The American Journal of Engineering and Technology
ISSN 2689-0984 | Open Access

TYPE Original Research
PAGE NO.212-217

Autonomous Cyber Defense through Reinforcement Learning and Simulation Environments: Foundations, Vulnerabilities, and Future Trajectories

Dr. Alexander M. Hartwell
Department of Computer Science, University of Edinburgh, United Kingdom

OPEN ACCESS

SUBMITED 01 July 2025
ACCEPTED 15 July 2025
PUBLISHED 31 July 2025
VOLUME Vol.07 Issue 07 2025

CITATION

Dr. Alexander M. Hartwell. (2025). Autonomous Cyber Defense through Reinforcement Learning and Simulation Environments: Foundations, Vulnerabilities, and Future Trajectories. The American Journal of Engineering and Technology, 7(07), 212–217. Retrieved from https://theamericanjournals.com/index.php/tajet/article/view/7095

Abstract: The accelerating complexity, scale, and adversarial nature of modern cyber environments have rendered traditional, human-centric cyber defense strategies increasingly insufficient. Autonomous cyber defense, particularly approaches grounded in reinforcement learning and simulation-based experimentation, has emerged as a promising paradigm capable of adapting to dynamic threats, reasoning under uncertainty, and responding at machine speed. This article presents a comprehensive, theory-driven research investigation into autonomous cyber defense systems, with a particular focus on reinforcement learning agents trained within cyber simulation environments. Drawing exclusively on the provided body of literature, the study synthesizes advances in cyber operations research gyms, autonomous agent design, reward shaping, adversarial robustness, and the emerging threat of poisoned or trojaned learning agents. The article methodologically integrates insights from foundational intrusion detection research, deep reinforcement learning, stochastic games, and graph-based reasoning to articulate a unified conceptual framework for autonomous cyber defense. Results are presented as an extensive descriptive analysis of observed patterns, theoretical behaviors, and empirical findings reported across prior studies, emphasizing both capabilities and vulnerabilities. The discussion critically

interrogates the limitations of current approaches, including closed-world assumptions, dataset bias, reward misalignment, and susceptibility to adversarial manipulation, while also exploring counter-arguments and mitigation strategies. Finally, the article outlines future research directions, emphasizing trustworthy autonomy, causal reasoning, and zero-day threat mitigation in complex software ecosystems. By providing a deeply elaborated, publication-ready synthesis, this work aims to serve as a foundational reference for researchers and practitioners seeking to advance the state of autonomous cyber defense.

Keywords: Autonomous cyber defense, reinforcement learning, cyber simulation, intrusion detection, adversarial machine learning, security agents

Introduction

Cybersecurity has undergone a profound transformation over the past two decades, driven by the exponential growth of networked systems, the proliferation of cloud and Internet of Things infrastructures, and the increasing sophistication of adversarial actors. Early cyber defense mechanisms were largely rule-based, reactive, and dependent on static signatures derived from known attack patterns. While effective against well-characterized threats, such approaches have struggled to cope with the volume, velocity, and variability of contemporary cyber attacks, particularly those involving zero-day exploits, polymorphic malware, and coordinated multi-stage intrusions (Sommer & Paxson, 2010; Buczak & Guven, 2016).

In response to these challenges, the cybersecurity research community has progressively embraced machine learning as a means of automating detection, classification, and response. Supervised and unsupervised learning methods have been extensively explored for intrusion detection, anomaly detection, and malware classification, yielding notable improvements in accuracy and scalability (Dalal & Rele, 2018; Ullah & Mahmoud, 2019; Rele & Patil, 2023). However, these methods remain fundamentally limited by their reliance on historical data, their sensitivity to dataset bias, and their inability to reason strategically over time. As highlighted by Sommer and Paxson (2010), many machine learning approaches implicitly assume a closed-world model that fails to reflect the adaptive and adversarial nature of real-world cyber environments.

Reinforcement learning has emerged as a compelling alternative paradigm, offering the capacity for agents to learn sequential decision-making policies through interaction with an environment. Unlike traditional machine learning models that passively analyze data, reinforcement learning agents actively explore their environment, receive feedback in the form of rewards, and iteratively refine their behavior. This makes reinforcement learning particularly well-suited to cyber defense tasks that involve continuous monitoring, adaptive response, and long-term optimization (Applebaum et al., 2022; Andrew et al., 2022).

The feasibility of reinforcement learning for cyber defense has been significantly advanced by the development of realistic cyber simulation environments. Cyber operations research gyms, such as CybORG, provide controlled yet expressive environments in which autonomous agents can be trained, evaluated, and compared under reproducible conditions (Baillie et al., 2020; Standen et al., 2022). These environments abstract complex cyber infrastructures into structured state spaces, action sets, and stochastic dynamics, enabling systematic experimentation while avoiding the ethical and operational risks of live-network testing.

Despite these advances, the deployment of autonomous cyber defense agents raises profound theoretical, technical, and ethical questions. Reinforcement learning agents are known to be sensitive to reward design, exploration strategies, and environmental assumptions, which can lead to unintended behaviors or brittle policies (Bates et al., 2023). Moreover, recent research has demonstrated that learning agents themselves can become targets of attack, through techniques such as reward poisoning, in-distribution triggers, and trojaned policies that activate malicious behaviors under specific conditions (Ashcraft & Karra, 2021; Acharya et al., 2023).

This article addresses these challenges by providing an exhaustive, theory-rich examination of autonomous cyber defense systems grounded in reinforcement learning and simulation. The primary contribution is not the introduction of new empirical experiments, but rather the integration and deep elaboration of existing findings into a coherent research narrative that identifies conceptual gaps, reconciles competing perspectives, and articulates future research trajectories. By synthesizing insights from cyber simulation, reinforcement learning theory, adversarial machine learning, and intrusion detection research, this work aims to advance the intellectual foundations of

autonomous cyber defense.

Methodology

The methodological approach of this research is grounded in qualitative synthesis, conceptual integration, and theoretical elaboration, drawing exclusively from the provided corpus of references. Rather than conducting new experimental evaluations, the study systematically analyzes the methodologies, assumptions, and findings reported across prior works to construct a unified understanding of autonomous cyber defense through reinforcement learning.

The first methodological pillar involves the examination of cyber simulation environments as experimental substrates for learning agents. CybORG and related cyber gyms are treated as methodological artifacts that shape the kinds of behaviors, policies, and vulnerabilities that agents can exhibit (Baillie et al., 2020; Standen et al., 2022). Particular attention is paid to how these environments model network topology, host configurations, attacker actions, defender observability, and stochastic outcomes. By analyzing these design choices, the study elucidates how simulation fidelity and abstraction influence the generalizability of learned policies.

The second pillar focuses on reinforcement learning paradigms employed in cyber defense research. This includes tabular methods, deep reinforcement learning, graph-based representations, and stochastic game formulations (Applebaum et al., 2022; Ammanabrolu & Riedl, 2018; Benaddi et al., 2022). The methodology involves a detailed comparison of how different learning paradigms encode state information, balance exploration and exploitation, and handle partial observability. Special emphasis is placed on causal reasoning and graph-based learning, which have been proposed as mechanisms for improving interpretability and robustness in complex cyber environments (Andrew et al., 2022; Shukla, 2025).

The third methodological component addresses adversarial considerations. The study systematically reviews research on adversarial attacks against learning agents, including reward manipulation, trigger-based poisoning, and universal trojan signatures (Ashcraft & Karra, 2021; Acharya et al., 2023). These works are analyzed not merely as isolated vulnerabilities, but as manifestations of deeper theoretical tensions between optimization objectives and security guarantees.

Finally, the methodology incorporates a critical engagement with intrusion detection literature to contextualize reinforcement learning within the broader history of machine learning for cybersecurity. Surveys and taxonomies of threats, datasets, and detection techniques are used to identify recurring challenges such as dataset bias, concept drift, and evaluation realism (Buczak & Guven, 2016; Hindy et al., 2020; Gao et al., 2020).

Through this multi-layered methodological synthesis, the article constructs a descriptive yet analytical account of autonomous cyber defense, emphasizing theoretical coherence, depth of explanation, and critical reflection.

Results

The synthesis of prior research reveals several recurring patterns and findings that collectively characterize the current state of autonomous cyber defense systems based on reinforcement learning. These results are presented descriptively, emphasizing conceptual insights rather than numerical metrics.

One prominent result is the demonstrated feasibility of training autonomous cyber defense agents within simulated environments. Studies utilizing CybORG and similar platforms consistently show that reinforcement learning agents can learn non-trivial defensive behaviors, such as identifying compromised hosts, prioritizing remediation actions, and managing limited defensive resources (Baillie et al., 2020; Standen et al., 2022). Even relatively simple tabular Q-learning agents have been shown to outperform static or heuristic-based baselines under certain conditions, particularly when the environment dynamics are sufficiently constrained (Applebaum et al., 2022).

Another significant finding concerns the role of representation in learning effectiveness. Graph-based state representations, inspired by work in text-adventure games and causal inference, enable agents to capture relational information about networks, processes, and privileges that would be difficult to encode in flat feature vectors (Ammanabrolu & Riedl, 2018; Andrew et al., 2022). These representations support more structured reasoning and appear to facilitate transfer across scenarios, suggesting a pathway toward more generalizable cyber defense policies.

Reward shaping emerges as a critical determinant of agent behavior. Bates et al. (2023) demonstrate that

poorly designed reward functions can lead to agents that optimize for superficial metrics while neglecting deeper security objectives, such as long-term system resilience or stealthy adversary detection. Conversely, carefully shaped rewards can produce agents that exhibit more human-aligned behaviors, such as proactive threat hunting and risk-aware decision-making.

The results also highlight the fragility of learning agents in adversarial settings. Multiple studies show that reinforcement learning agents can be manipulated through subtle perturbations to their training environment or reward signals. Ashcraft and Karra (2021) demonstrate that in-distribution triggers can be embedded in training data, causing agents to behave maliciously when specific conditions are met. Acharya et al. (2023) further reveal the existence of universal trojan signatures that can compromise a wide range of reinforcement learning policies, raising concerns about the trustworthiness of autonomous agents deployed in security-critical contexts.

From the intrusion detection perspective, the results reaffirm longstanding concerns about dataset representativeness and closed-world assumptions. Despite advances in deep learning and reinforcement learning, many systems continue to rely on benchmark datasets that fail to capture the diversity and evolution of real-world threats (Hindy et al., 2020; Sommer & Paxson, 2010). This limitation extends to simulation environments, which, while more flexible, still encode implicit assumptions that may not hold outside the laboratory.

Collectively, these results suggest that autonomous cyber defense is both promising and perilous. Reinforcement learning agents can achieve levels of adaptability and responsiveness unattainable by traditional systems, but they also introduce new attack surfaces and epistemic uncertainties.

Discussion

The findings synthesized in this article invite a nuanced discussion that balances optimism about autonomous cyber defense with a sober assessment of its limitations and risks. At a theoretical level, reinforcement learning offers a powerful framework for modeling cyber defense as a sequential decision-making problem under uncertainty. This framing aligns well with the realities of cyber operations, where defenders must continuously allocate attention and resources in the face of incomplete information and adaptive adversaries (Andrew et al., 2022).

However, the reliance on simulation environments raises fundamental questions about validity and transferability. While cyber gyms like CybORG represent a significant methodological advance, they inevitably simplify reality. Network configurations, attacker behaviors, and defensive actions are discretized and bounded, potentially biasing learned policies toward artifacts of the simulation rather than robust strategies. Critics may argue that such environments risk creating a false sense of progress, analogous to early successes in game-playing AI that failed to translate to real-world complexity.

Counter-arguments emphasize that all empirical science relies on models and abstractions, and that simulation-based research provides a necessary stepping stone toward real-world deployment. The key challenge, therefore, is not to eliminate abstraction, but to make it explicit, diverse, and continually evolving (Baillie et al., 2020). Incorporating stochasticity, partial observability, and heterogeneous attacker models can mitigate some of the risks of overfitting to simplified environments.

The vulnerability of reinforcement learning agents to adversarial manipulation represents another critical concern. The discovery of trojaned policies and poisoned reward signals undermines the assumption that learning agents inherently enhance security. From a defensive standpoint, this creates a paradox: systems designed to protect against attackers may themselves become vectors of compromise. Addressing this paradox requires rethinking trust models, verification techniques, and the role of human oversight in autonomous systems (Acharya et al., 2023; Ashcraft & Karra, 2021).

Reward shaping, while essential for guiding learning, also introduces normative assumptions about what constitutes desirable behavior. Security objectives are inherently multi-dimensional and context-dependent, encompassing confidentiality, integrity, availability, and resilience. Encoding these objectives into scalar reward functions risks oversimplification and unintended consequences (Bates et al., 2023). Future research may benefit from multi-objective reinforcement learning or hierarchical approaches that better reflect the layered nature of security goals.

From the broader intrusion detection literature, the persistence of closed-world assumptions serves as a

cautionary tale. Even the most sophisticated learning algorithms cannot compensate for blind spots in data and modeling. Autonomous cyber defense systems must therefore be designed with humility, recognizing the limits of their knowledge and the inevitability of surprise (Sommer & Paxson, 2010).

Looking forward, several promising directions emerge. Causal reasoning and graph-based representations offer pathways toward more interpretable and robust agents, capable of reasoning about cause and effect rather than merely correlational patterns (Shukla, 2025). Integrating reinforcement learning with stochastic game theory may enable more realistic modeling of attacker-defender dynamics, capturing strategic interactions over time (Benaddi et al., 2022). Finally, advances in verification and adversarial robustness could help ensure that autonomous agents behave safely even under malicious influence (Gao et al., 2020).

Conclusion

Autonomous cyber defense, enabled by reinforcement learning and cyber simulation environments, represents a transformative yet challenging frontier in cybersecurity research. The body of literature examined in this article demonstrates both the potential and the perils of delegating defensive decision-making to learning agents. On the one hand, such agents can adapt to dynamic threats, operate at machine speed, and uncover strategies beyond human intuition. On the other hand, they introduce new vulnerabilities, ethical dilemmas, and epistemic uncertainties that demand careful consideration.

By synthesizing insights from cyber operations research gyms, reinforcement learning theory, intrusion detection, and adversarial machine learning, this article has articulated a comprehensive conceptual framework for understanding autonomous cyber defense. The analysis underscores the importance of simulation fidelity, representation learning, reward design, and adversarial robustness, while also highlighting enduring challenges related to generalization and trust.

Ultimately, the path forward lies not in uncritical enthusiasm nor in outright skepticism, but in rigorous, theory-informed research that acknowledges complexity and embraces interdisciplinary perspectives. Autonomous cyber defense should be viewed as a complement to, rather than a replacement for, human expertise, embedded within socio-technical systems that prioritize resilience, transparency, and accountability. As cyber threats continue to evolve, so too must our approaches to defense, guided by both technological innovation and critical reflection.

References

1. Acharya, M., Zhou, W., Roy, A., Lin, X., Li, W., & Jha, S. (2023). Universal trojan signatures in reinforcement learning. Proceedings of the NeurIPS 2023 Workshop on Backdoors in Deep Learning: The Good, the Bad, and the Ugly.
2. Ammanabrolu, P., & Riedl, M. O. (2018). Playing text-adventure games with graph-based deep reinforcement learning. arXiv preprint arXiv:1812.01628.
3. Andrew, A., Spillard, S., Collyer, J., & Dhir, N. (2022). Developing optimal causal cyber-defence agents via cyber security simulation. arXiv preprint arXiv:2207.12355.
4. Applebaum, A., Dennler, C., Dwyer, P., Moskowitz, M., Nguyen, H., Nichols, N., Park, N., Rachwalski, P., Rau, F., Webster, A., & Wolk, M. (2022). Bridging automated to autonomous cyber defense: Foundational analysis of tabular q-learning. Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security.
5. Ashcraft, C., & Karra, K. (2021). Poisoning deep reinforcement learning agents with in-distribution triggers. arXiv preprint arXiv:2106.07798.
6. Baillie, C., Standen, M., Schwartz, J., Docking, M., Bowman, D., & Kim, J. (2020). CybORG: An autonomous cyber operations research gym. arXiv preprint arXiv:2002.10667.
7. Bates, E., Mavroudis, V., & Hicks, C. (2023). Reward shaping for happier autonomous cyber security agents. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security.
8. Benaddi, H., Elhajji, S., Benaddi, A., Amzazi, S., & Oudani, H. (2022). Robust enhancement of intrusion detection systems using deep reinforcement learning and stochastic game. IEEE Transactions on Vehicular Technology, 71(10), 11089–11102.
9. Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176.
10. Dalal, K. R., & Rele, M. (2018). Cyber security: Threat

detection model based on machine learning algorithm. Proceedings of the 3rd International Conference on Communication and Electronics Systems.

11. Defense Advanced Research Projects Agency. (2023). Cyber agents for security testing and learning environments. Retrieved from https://sam.gov.

12. Gao, J., Korolov, R., & Kantarcioglu, M. (2020). Adversarial attacks and defenses for deep learning-based network intrusion detection systems. Proceedings of the Annual Computer Security Applications Conference.

13. Hindy, H., Brosset, D., Bayne, E., Seeam, A., Tachtatzis, C., & Atkinson, R. (2020). A taxonomy of network threats and the effect of current datasets on intrusion detection systems. IEEE Access, 8, 104650–104675.

14. Queensland Defence Science Alliance. (2022). Artificial Intelligence for Decision Making Initiative. Retrieved from https://queenslanddefencesciencealliance.com.au.

15. Rele, M., & Patil, D. (2023). Intrusive detection techniques utilizing machine learning, deep learning, and anomaly-based approaches. Proceedings of the IEEE International Conference on Cryptography, Informatics, and Cybersecurity.

16. Shukla, O. (2025). Autonomous cyber defence in complex software ecosystems: A graph-based and AI-driven approach to zero-day threat mitigation. Journal of Emerging Technologies and Innovation Management, 1(01), 01–10.

17. Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. IEEE Symposium on Security and Privacy.

18. Standen, M., Bowman, D., Hoang, S., Richer, T., Lucas, M., Van Tassel, R., Vu, P., Kiely, M., Konschnik, N., & Collyer, J. (2022). Cyber operations research gym. Retrieved from https://github.com/cage-challenge/CybORG.

19. Ullah, I., & Mahmoud, Q. H. (2019). A two-level hybrid model for anomaly-based intrusion detection in IoT networks. Electronics, 8(12), 1396.