COLOMBO SCIENTIFIC PUBLISHING

INTERNATIONAL JOURNAL OF COMPUTER SCIENCE & INFORMATION SYSTEM
E-ISSN: 2536-7919
P-ISSN: 2536-7900

PAGE NO: 18-23

Causality-Aware and Explainable Reinforcement Learning Architectures for Autonomous Cyber Defense in Complex Networked Environments

Dr. Alexander M. Reinhardt
Department of Computer Science and Systems Engineering, University of Utrecht, Netherlands

ARTICLE INFO

Article history:
Submission: July 01, 2025
Accepted: July 17, 2025
Published: July 31, 2025
VOLUME: Vol.10 Issue 07 2025

Keywords:
Autonomous cyber defense, reinforcement learning, explainable AI, hierarchical learning, cyber resilience, adversarial environments

ABSTRACT

The accelerating scale, complexity, and adversarial sophistication of cyber threats have fundamentally challenged traditional, rule-based security mechanisms. Autonomous cyber defense has emerged as a promising paradigm capable of responding to attacks at machine speed, adapting to unknown threats, and operating continuously within dynamic network environments. Reinforcement learning has been positioned at the center of this paradigm due to its ability to learn optimal policies through interaction with an environment. However, deploying reinforcement learning in real-world cyber defense raises profound challenges related to robustness, explainability, causality, scalability, and adversarial manipulation. This article presents a comprehensive and theoretically grounded research synthesis that integrates hierarchical reinforcement learning, deception-based defense strategies, causality-driven policy discovery, and explainable artificial intelligence into a unified conceptual framework for autonomous cyber defense. Drawing strictly from established academic literature and simulation platforms, the study examines how reinforcement learning agents can be trained using realistic cyber ranges and attack simulators to generalize across novel threat scenarios while maintaining resilience against adversarial interference. Particular emphasis is placed on optimal stopping formulations for intrusion prevention, feedback-enabled cyber resilience, and graph-based representations of attack–defense interactions. The article advances the argument that next-generation autonomous defense systems must move beyond performance-centric optimization toward architectures that explicitly encode causal reasoning, strategic deception, and human-interpretable decision processes. Through extensive methodological elaboration and interpretive analysis, the work identifies key research gaps and outlines future directions for building trustworthy, adaptive, and operationally viable autonomous cyber defense systems capable of functioning within complex software ecosystems and critical infrastructure networks.

INTRODUCTION

The contemporary cyber threat landscape is characterized by unprecedented complexity, persistence, and strategic sophistication. Modern attackers leverage automation, artificial intelligence, and deep domain knowledge to exploit vulnerabilities across vast and heterogeneous digital ecosystems. Advanced persistent threats, zero-day exploits, and coordinated multi-stage attacks have rendered static and manually configured security controls increasingly inadequate. Traditional defensive mechanisms, such as signature-based intrusion detection systems or predefined access control policies, struggle to adapt to rapidly evolving adversarial behaviors and often fail to provide timely responses at the scale required by modern networks (Ridley, 2018; Vyas et al., 2023).

In response to these challenges, the concept of autonomous cyber defense has gained significant traction within both academic and applied research communities. Autonomous cyber defense refers to the ability of computational agents to monitor, detect, reason about, and respond to cyber threats with minimal or no

human intervention. At its core, this paradigm seeks to emulate certain aspects of human defensive reasoning, such as situational awareness, learning from experience, and strategic adaptation, while surpassing human limitations in speed, consistency, and scalability (Dhir et al., 2021; Molina-Markham et al., 2021).

Reinforcement learning has emerged as a particularly compelling foundation for autonomous cyber defense systems. Reinforcement learning enables an agent to learn a policy that maximizes long-term reward through trial-and-error interactions with an environment. This learning paradigm is well-suited to adversarial domains where explicit models of attacker behavior are incomplete or unavailable, and where optimal defensive actions depend on sequential decision-making under uncertainty (Mnih et al., 2013; Huang and Zhu, 2021). However, the direct application of reinforcement learning to cyber defense introduces a host of unresolved issues that limit its practical adoption.

One of the most pressing challenges is the problem of explainability. Reinforcement learning agents, particularly those based on deep neural networks, often operate as opaque black boxes. While such agents may achieve high performance in simulation, their internal decision-making processes are difficult to interpret or justify. In high-stakes security contexts, where defensive actions can disrupt critical services or escalate conflicts, the inability to explain why a particular action was taken undermines trust, accountability, and regulatory compliance (Hoffman et al., 2018; Hamon et al., 2020).

Another critical concern relates to robustness and adversarial manipulation. Cyber environments are inherently adversarial, and attackers may actively seek to deceive or manipulate learning agents by altering observations, rewards, or cost signals. Research has demonstrated that reinforcement learning agents are vulnerable to adversarial perturbations that can significantly degrade performance or induce harmful behaviors (Huang and Zhu, 2019). This vulnerability raises fundamental questions about the reliability of autonomous defense systems deployed in contested environments.

Scalability and generalization also pose substantial obstacles. Real-world networks are large, heterogeneous, and continuously evolving. Training reinforcement learning agents that can generalize across diverse network topologies, attack strategies, and operational constraints remains an open research problem. While simulation environments such as CyberBattleSim and NASimEmu provide valuable platforms for experimentation, bridging the gap between simulated training and real-world deployment requires careful consideration of abstraction, transfer learning, and domain randomization (Microsoft Defender Research Team, 2021; Janisch et al., 2023).

The literature further highlights the need for strategic reasoning beyond reactive defense. Effective cyber defense often involves proactive measures such as deception, moving target defense, and anticipatory resource allocation. Incorporating these strategies into reinforcement learning frameworks necessitates richer representations of attacker–defender interactions, including game-theoretic models and hierarchical decision structures (Hofer et al., 2019; Kordy et al., 2010).

Despite significant progress across these individual research threads, there remains a notable lack of integrative frameworks that address explainability, robustness, causality, and strategic depth in a unified manner. Existing studies tend to focus on isolated aspects of the problem, such as learning optimal stopping policies for intrusion prevention (Hammar and Stadler, 2021) or applying hierarchical reinforcement learning to abstract decision-making (Hengst, 2010; Hu et al., 2022). What is missing is a comprehensive synthesis that articulates how these elements can be coherently combined to support autonomous cyber defense in complex, real-world environments.

This article seeks to address this gap by presenting an original, publication-ready research synthesis that draws strictly on established academic references to develop a causality-aware and explainable reinforcement learning architecture for autonomous cyber defense. Rather than proposing a specific algorithmic implementation, the work emphasizes theoretical integration, methodological rigor, and interpretive depth. By elaborating on the implications, limitations, and future directions of existing approaches, the article aims to provide a conceptual foundation for researchers and practitioners seeking to design trustworthy and resilient autonomous defense systems.

Methodology

The methodological approach adopted in this research is grounded in analytical synthesis rather than empirical experimentation. Given the complexity and interdisciplinarity of autonomous cyber defense, a

purely experimental methodology would risk obscuring foundational theoretical issues. Instead, the methodology focuses on systematically integrating insights from reinforcement learning theory, cyber security modeling, game theory, explainable artificial intelligence, and adversarial robustness. This integrative approach enables a deep examination of how diverse methodological strands can be coherently aligned to address the multifaceted challenges of autonomous cyber defense.

At the core of the methodological framework is the reinforcement learning paradigm, which conceptualizes cyber defense as a sequential decision-making problem under uncertainty. In this formulation, the defender is modeled as an agent interacting with an environment that represents the network, its assets, and potential adversarial actions. The agent observes the state of the environment, selects defensive actions, and receives feedback in the form of rewards or penalties that reflect security outcomes such as attack mitigation, service availability, and operational cost (Huang and Zhu, 2021).

However, traditional flat reinforcement learning formulations struggle to scale in high-dimensional cyber environments. To address this issue, the methodology incorporates hierarchical reinforcement learning as a foundational design principle. Hierarchical reinforcement learning decomposes complex decision-making tasks into multiple levels of abstraction, enabling agents to learn high-level strategies and low-level tactics concurrently (Hengst, 2010). In the context of cyber defense, high-level policies may govern strategic objectives such as prioritizing critical assets or deploying deception, while low-level policies handle specific actions such as patching vulnerabilities or isolating compromised hosts.

The hierarchical structure is further informed by causality-driven discovery mechanisms. Recent advances in reinforcement learning have demonstrated the value of explicitly modeling causal relationships between actions, environmental states, and outcomes. By identifying causal hierarchies, agents can learn more robust and transferable policies that generalize across different scenarios (Hu et al., 2022). Methodologically, this involves structuring the state and action spaces in ways that reflect underlying causal mechanisms, such as the propagation of attacks through network graphs or the delayed effects of defensive interventions.

To ensure realism and relevance, the methodological framework leverages simulation-based training environments that emulate realistic cyberattack scenarios. Platforms such as Cyber BattleSim and NASimEmu provide controlled yet flexible environments in which agents can be exposed to diverse attack strategies and network configurations (Microsoft Defender Research Team, 2021; Janisch et al., 2023). These simulators are particularly valuable for studying generalization, as they allow systematic variation of environmental parameters and attacker behaviors.

A critical methodological consideration is the incorporation of adversarial manipulation and deception. Cyber defenders do not operate in neutral environments; attackers actively seek to mislead and exploit defensive mechanisms. The methodology therefore integrates models of deceptive reinforcement learning, in which reward signals or observations may be intentionally distorted by adversaries (Huang and Zhu, 2019). Studying agent behavior under such conditions provides insights into robustness and the limits of learning-based defense.

Explain ability is addressed through the adoption of metrics and evaluation criteria derived from explainable AI research. Rather than treating explain ability as an afterthought, the methodology embeds interpretability considerations into the design of learning architectures and policy representations. This includes the use of hierarchical and graph-based models that naturally lend themselves to human interpretation, as well as the application of established explain ability metrics to assess the transparency and comprehensibility of agent decisions (Hoffman et al., 2018; Hamon et al., 2020).

Finally, the methodology incorporates game-theoretic reasoning to model the strategic interaction between attackers and defenders. Attack–defense trees and their equivalence to extensive-form games provide a formal framework for analyzing strategic choices and potential equilibria (Jhawar et al., 2016; Kordy et al., 2010). By aligning reinforcement learning objectives with game-theoretic insights, the methodology seeks to ensure that learned policies are not only effective in simulation but also strategically sound in adversarial contexts.

Results

The analytical synthesis conducted in this research yields several significant findings regarding the design and potential performance of autonomous cyber defense systems based on reinforcement learning. While

the results are not derived from numerical experimentation, they emerge from a detailed comparison and integration of existing theoretical and applied studies.

One of the most salient findings is that hierarchical reinforcement learning substantially enhances the scalability and interpretability of autonomous defense agents. By decomposing decision-making into multiple levels, hierarchical architectures reduce the cognitive and computational burden associated with large action spaces. This decomposition aligns closely with how human defenders conceptualize security operations, distinguishing between strategic planning and tactical execution. As a result, hierarchical agents are better positioned to manage complex networks and adapt to evolving threats without requiring exhaustive retraining (Hengst, 2010; Hu et al., 2022).

The synthesis also reveals that causality-aware learning architectures offer significant advantages in terms of robustness and generalization. Agents that explicitly model causal relationships are less likely to overfit to spurious correlations present in training environments. This capability is particularly important in cyber defense, where superficial patterns may change rapidly as attackers adapt their tactics. By grounding policy learning in causal structures, agents can transfer learned behaviors to novel scenarios more effectively, as demonstrated in studies involving network attack simulators (Janisch et al., 2023; Nyberg and Johnson, 2023).

Another important result concerns the role of deception in autonomous defense. The integration of deception-based strategies, such as honeypots and decoy systems, into reinforcement learning frameworks enables agents to actively shape attacker behavior rather than merely reacting to it. Model-driven deception has been shown to increase uncertainty for attackers and delay or disrupt attack progression (Hofer et al., 2019). When combined with reinforcement learning, deception becomes an adaptive tool that can be optimized over time based on observed adversarial responses.

The analysis further highlights the critical importance of optimal stopping formulations for intrusion prevention. Rather than continuously applying defensive actions, agents must decide when to intervene to maximize long-term security outcomes while minimizing operational costs. Optimal stopping models provide a principled approach to balancing these trade-offs, enabling agents to defer action until sufficient evidence of malicious activity has accumulated (Hammar and Stadler, 2021). This finding underscores the need for temporal reasoning and patience in autonomous defense policies.

Explainability emerges as both a challenge and an opportunity. The synthesis indicates that while deep reinforcement learning models are inherently complex, their integration with hierarchical and graph-based representations can significantly enhance interpretability. Explainable AI metrics provide a means of systematically evaluating how well agents' decisions can be understood by human operators. This capability is essential for building trust and facilitating human–machine collaboration in security operations (Hoffman et al., 2018; Hamon et al., 2020).

Finally, the results suggest that simulation-based training environments play a crucial role in advancing autonomous cyber defense research. Platforms such as CyberBattleSim and NASimEmu enable systematic exploration of defensive strategies and facilitate reproducibility. However, the synthesis also cautions that simulation fidelity and scenario diversity are critical determinants of real-world applicability. Overreliance on simplified environments risks producing agents that perform well in simulation but fail under real-world conditions (Microsoft Defender Research Team, 2021; Janisch et al., 2023).

Discussion

The findings of this research synthesis carry significant implications for the future of autonomous cyber defense. By integrating hierarchical reinforcement learning, causality-driven reasoning, deception strategies, and explainable AI, the proposed conceptual framework addresses many of the limitations identified in existing approaches. However, the discussion also reveals important tensions, trade-offs, and unresolved challenges that warrant careful consideration.

One of the central tensions concerns the balance between autonomy and human oversight. While autonomous agents offer the promise of rapid and adaptive defense, complete autonomy may be neither feasible nor desirable in high-stakes environments. Explainability and interpretability serve as critical bridges between machine autonomy and human judgment. By enabling operators to understand and anticipate agent behavior, explainable reinforcement learning architectures support collaborative defense

models in which humans retain strategic control while delegating tactical decisions to machines (Hoffman et al., 2018).

Another key issue relates to adversarial robustness. The analysis underscores that reinforcement learning agents are inherently vulnerable to manipulation, particularly in environments where attackers can influence observations or rewards. While causality-aware learning and deception strategies enhance robustness, they do not eliminate the risk of exploitation. Future research must therefore explore additional safeguards, such as adversarial training, uncertainty estimation, and hybrid architectures that combine learning-based and rule-based components (Huang and Zhu, 2019; Hamon et al., 2020).

The discussion also highlights the importance of strategic depth in autonomous defense. Reactive defense alone is insufficient against sophisticated adversaries. By incorporating game-theoretic reasoning and attack–defense trees, reinforcement learning agents can anticipate attacker strategies and allocate resources more effectively. However, modeling strategic interactions at scale remains computationally challenging, and approximations are often necessary. Understanding the limits of these approximations is crucial for avoiding overconfidence in learned policies (Jhawar et al., 2016; Kordy et al., 2010).

Limitations of the current synthesis must be acknowledged. The reliance on existing literature means that empirical validation of the integrated framework remains an open task. Moreover, while simulation environments provide valuable insights, they cannot fully capture the socio-technical complexities of real-world cyber operations. Issues such as organizational constraints, legal considerations, and human behavior fall outside the scope of most technical models but significantly influence defense outcomes (Vyas et al., 2023).

Looking ahead, several promising directions for future research emerge. One avenue involves the development of standardized benchmarks and evaluation protocols for autonomous cyber defense systems. Such benchmarks would facilitate systematic comparison of approaches and accelerate progress. Another direction concerns the integration of multi-agent reinforcement learning, in which multiple defensive agents coordinate their actions across distributed networks. Cooperative decision-making frameworks have shown potential for enhancing resilience and scalability (Cheah et al., 2023).

Finally, there is a pressing need for interdisciplinary collaboration. Autonomous cyber defense sits at the intersection of computer science, security engineering, cognitive science, and policy. Advancing the field will require not only technical innovation but also thoughtful engagement with ethical, legal, and societal considerations.

Conclusion

This article has presented a comprehensive and theoretically grounded synthesis of reinforcement learning-based approaches to autonomous cyber defense. By drawing strictly on established academic literature, the work has articulated a unified framework that integrates hierarchical learning, causality-driven reasoning, deception strategies, explainable AI, and game-theoretic modeling. The analysis demonstrates that addressing the challenges of scalability, robustness, and trustworthiness requires moving beyond isolated technical solutions toward holistic architectures that reflect the complexity of real-world cyber environments.

The findings underscore that autonomous cyber defense is not merely a problem of optimization but a fundamentally strategic and socio-technical endeavor. Reinforcement learning provides powerful tools for adaptation and learning, but its effective deployment depends on careful architectural design and rigorous evaluation. By emphasizing explainability and causal reasoning, the proposed framework aligns machine decision-making more closely with human understanding and oversight.

As cyber threats continue to evolve, the need for adaptive and trustworthy defense mechanisms will only intensify. This research contributes to the ongoing effort to conceptualize and design autonomous cyber defense systems that are not only effective but also interpretable, resilient, and aligned with human values. In doing so, it lays the groundwork for future empirical studies and practical implementations that can bring the promise of autonomous defense closer to operational reality.

References

1. Bridges, R. A., Rice, A. E., Oesch, S., Nichols, J. A., Watson, C., Spakes, K., Norem, S., Huettel, M., Jewell, B., Weber, B., et al. (2023). Testing SOAR tools in use. Computers & Security, 129, 103201.
2. Cheah, M., Stone, J., Haubrick, P., Bailey, S., Rimmer, D., Till, D., Lacey, M., Kruczynska, J., Dorn, M. (2023). Co-decyber: Co-operative decision making for cybersecurity using deep multi-agent reinforcement learning. European Symposium on Research in Computer Security.
3. Dhir, N., Hoeltgebaum, H., Adams, N., Briers, M., Burke, A., Jones, P. (2021). Prospective artificial intelligence approaches for active cyber defence. arXiv preprint arXiv:2104.09981.
4. Hammar, K., Stadler, R. (2021). Learning intrusion prevention policies through optimal stopping. Proceedings of the International Conference on Network and Service Management.
5. Hamon, R., Junklewitz, H., Sanchez, I., et al. (2020). Robustness and explainability of artificial intelligence. Publications Office of the European Union.
6. Hengst, B. (2010). Hierarchical reinforcement learning. Springer US.
7. Hicks, C., Mavroudis, V., Foley, M., Davies, T., Highnam, K., Watson, T. (2023). Canaries and whistles: Resilient drone communication networks with (or without) deep reinforcement learning. ACM Workshop on Artificial Intelligence and Security.
8. Hofer, W., Edgar, T., Vrabie, D., Nowak, K. (2019). Model-driven deception for control system environments. IEEE International Symposium on Technologies for Homeland Security.
9. Hoffman, R. R., Mueller, S. T., Klein, G., Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.
10. Hu, X., Zhang, R., Tang, K., Guo, J., Yi, Q., Chen, R., Du, Z., Li, L., Guo, Q., Chen, Y., et al. (2022). Causality-driven hierarchical structure discovery for reinforcement learning. Advances in Neural Information Processing Systems, 35.
11. Huang, L., Zhu, Q. (2019). Adaptive strategic cyber defense for advanced persistent threats in critical infrastructure networks. SIGMETRICS Performance Evaluation Review, 46(2), 52–56.
12. Huang, Y., Zhu, Q. (2019). Deceptive reinforcement learning under adversarial manipulations on cost signals. GameSec Proceedings.
13. Huang, Y., Huang, L., Zhu, Q. (2021). Reinforcement learning for feedback-enabled cyber resilience. Annual Reviews in Control, 53, 273–295.
14. Janisch, J., Pevny, T., Lisý, V. (2023). NASimEmu: Network attack simulator & emulator for training agents generalizing to novel scenarios. European Symposium on Research in Computer Security.
15. Jhawar, R., Mauw, S., Zakiuddin, I. (2016). Automating cyber defence responses using attack-defence trees and game theory. European Conference on Cyber Warfare and Security.
16. Kordy, B., Mauw, S., Melissen, M., Schweitzer, P. (2010). Attack–defense trees and two-player binary zero-sum extensive form games are equivalent. GameSec Proceedings.
17. Microsoft Defender Research Team. (2021). CyberBattleSim.
18. Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. arXiv preprint arXiv:1312.5602.
19. Molina-Markham, A., Miniter, C., Powell, B., Ridley, A. (2021). Network environment design for autonomous cyberdefense. arXiv preprint arXiv:2103.07583.
20. Nyberg, J., Johnson, P. (2023). Training automated defense strategies using graph-based cyber attack simulations. arXiv preprint arXiv:2304.11084.
21. Ridley, A. (2018). Machine learning for autonomous cyber defense. The Next Wave, 22(1), 7–14.
22. 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.
23. Vyas, S., Hannay, J., Bolton, A., Burnap, P. (2023). Automated cyber defence: A review. arXiv preprint arXiv:2303.04926.