When AI Meets AI: A Game-Theoretic Defense Framework Against AI Empowered Cyber Threats

Xinyu Li

Cyberspace Institute of Advanced Technology
Guangzhou University, China
XinyuLi9@e.gzhu.edu.cn

Abstract

The widespread adoption of artificial intelligence (AI) in cybersecurity has led to the emerging threat of AI-driven cyberattacks, such as LLM-empowered Advanced Persistent Threats (APTs), challenging the effect of conventional deception defense mechanisms. To fill this critical gap, my work aims to develop a game-theoretic defense AI agent capable of providing the optimal deception resource deployment strategy, to establish AI-driven defenses against AI-empowered cyberattacks. In this proposal, I model the attacker and defender interaction as a dynamic game with incomplete information between AI agents, and then derive the equilibrium defense strategies. Synthetic data based experiments and real-world implementations would be conducted to validate the proposed framework. This study has the potential to improve the effectiveness of deception defense in three dimensions: scalability, real-time capability, and strategic intelligence.

Introduction

Deception defenses (DD) is a traditional proactive technology to confront cyber threats, through the deployment of decoy systems such as honeypots to waste attackers' resources and reveal their attacking tactic, techniques and procedures (TTPs) with a low rate of false detections (Javadpour et al. 2024). However, the effectiveness of traditional DD methods is significantly discounted when coping with intelligent attacks, such as Advanced Persistent Threats (APTs). (Zhu et al. 2021). The defensive mission is critically challenged by a fundamental asymmetry: an attacker needs to succeed only once, whereas a defender must be successful every time. This imbalance is further amplified by the rise of AI-powered attacks, which lower the barrier to one-time success (Lohn 2025). This inherent disadvantage is further exacerbated by the static nature of existing DD, which lack the dynamic adaptability to counter AI-driven attacks that continuously evolve based on reconnaissance (Hausken, Welburn, and Zhuang 2024).

In this study, I propose a game-theoretic AI defense framework to enhance DD decision-making intelligence, where defenders and attackers are formalized as strategic agents empowered by AI. We consider that for the evolving landscape of AI-driven cybersecurity, a promising solution is to address these challenges by deploying AI against AI itself.

Background

Since being first proposed by Gene Spafford in 1989, DD has seen significant research advancements (Beltrán-López, Pérez, and Nespoli 2025).

In the integration of DD and game theory (GT), some representative works have explored strategic interactions. In (Pawlick, Colbert, and Zhu 2019), a GT-based taxonomy was proposed, modeling how defenders and attackers interact. Anwar et al. developed a Partially Observable Stochastic Game (POSG) model (Anwar, Kamhoua, and Leslie 2020). In this model, defenders introduce honeypots as their strategy, while attackers make decisions based on limited knowledge of the state of the network. This allows real-time simulation of how both sides' choices affect network security. Florea et al. also explored a GT approach for network security with honeypots (Florea and Craus 2022). It modeled potential attacks and adjusted defenses as a Stackelberg game, highlighting the value of informed and adaptable defensive strategies.

However, with the rise of AI, cyberattacks have become more and more intelligent (Guembe et al. 2022). For instance, large language models (LLMs) can generate phishing content or probe networks autonomously (Das, Amini, and Wu 2025), which reduces the cost of attacks for attackers, expands the scope of attacks, and improves the sophistication of attacks. Unfortunately, existing research on attacker modeling remains confined to the manual level, rendering them inadequate for novel AI-driven cyberattacks, which makes me think of using AI to counter AI.

Prior Work by the Applicant

In (Li et al. 2025), we successfully formalized sandwich attacks on blockchain into mathematical formulas and categorized defense strategies based on transaction phases, which is equally applicable to formalizing attacks targeting AI agents. In addition, I have ever developed an AI system with conventional neural networks to detect malicious images, which has laid a solid foundation for understanding AI fundamentals and its application in security.

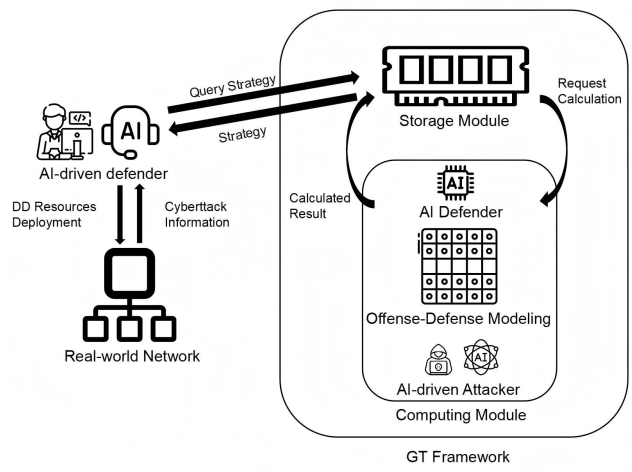

Figure 1: The proposed game-theoretic AI defense framework against AI-empowered attackers. The AI-driven defender first requests a strategy from the GT framework based on collected attack information. The GT framework first checks the Storage Module (SM). If a pre-computed corresponding strategy exists, it is directly provided to the human expert; otherwise, the Computing Module computes the current optimal strategy, which is then stored in the SM and provided to the human expert. The human expert ultimately makes the final decision and deploys DD resources.

Approach

Our proposed game-theoretic AI defense framework against AI-empowered attackers is presented in Figure 1. The game model between the defender and the attacker is formalized as $G(N, A, T, O, P, R)$.

Players We use the set $N = \{AI_D, AI_A\}$ to represent AI-driven players, where AI_D represents the defender and AI_A represents the attacker. Both are autonomous adaptive entities with cognitive capabilities shaped by AI models (e.g., LLMs) fine-tuned on cybersecurity datasets.

Action Space The action space is defined as $A = A_{AI_D} \times A_{AI_A}$, with A_{AI_D} focusing on the deployment of DD resources and adaptive adjustment, A_{AI_A} reflecting AI-driven attack behaviors that adapt to perceived environmental cues. After observing AI_D's action $a^t_{AI_D}$, AI_A selects $a^t_{AI_A} \in A_{AI_A}$, and their choices determine the reward values for the two players at time t, after which the next step of the game proceeds until one player terminates their actions.

Type Space $T = T_{AI_D} \times T_{AI_A}$ is the type space, where the type of each player encapsulates the private, stable attributes that affect their action preferences and payoff functions. For example, AI_D's resource budget type can be classified as $T^{budget}_{AI_D} = \{High, Medium, Low\}$.

Observation Space The observation space $O = O_{AI_D} \times O_{AI_A}$ models incomplete and noisy information that each AI agent receives about the state of the game, mimicking uncertainty in cyber environments. Observations are time-dependent, with $o^t_{AI_D} \in O_{AI_D}$ and $o^t_{AI_A} \in O_{AI_A}$ captured at each step t.

Probability Distributions The probability space P models probability in type prior (P_{type}), observation noise (P_{obs}), and state transitions (P_{trans}). Both agents update their beliefs using Bayesian inference, e.g., AI_D revises $P_{type}(t_{AI_A} = Advanced)$ upward if $o^t_{AI_D}$ shows AI_A avoided a low-fidelity honeypot.

Payoff The payoff function $R = \{R_{AI_D}, R_{AI_A}\}$ quantifies the utility of each AI agent. Defender payoff ($R_{AI_D} : T \times A \times O \rightarrow \mathbb{R}$) balances attack detection gains, resource costs, and time consumed; Attacker Payoff ($R_{AI_A} : T \times A \times O \rightarrow \mathbb{R}$) reflects attack progress minus the cost of avoiding deception. The specific formula will be determined once the cognitive model is finalized.

Evaluation and Discussion

We will conduct comparative experiments where our AI-driven DD method (experimental group) and traditional methods (control group) will face cyberattacks armed with AI highly likely to occur in real-world scenarios, ensuring the framework remains effective in practical missions. Core metrics include degree of TTPs detected (Sun et al. 2021), percentage of successful defenses, cost of deployed resources and response time, etc. This framework is expected to significantly enhance the effectiveness of DD for practical applications by:

- Higher strategic intelligence by leveraging AI vs. AI interactions to bridge the gap with sophisticated AI-driven cyberattacks;
- Better real-time responsiveness through pre-computed strategy storage that reduces redundant computations;
- Potential scalability due to its lightweight design, which enables seamless integration across networks of varying scales to support critical tasks.

Conclusion

This research addresses the challenge of AI powered cyberattacks by developing a game-theoretic AI defense framework. Unlike static methods, the approach constructs attacker–defender interactions as game models with incomplete information, enabling adaptive strategies through cognitive modeling and equilibrium analysis. Concurrently, it accelerates real-time responses by storing pre-computed strategy combinations. Simulation experiments and real-world validations will be conducted to demonstrate its performance over traditional defenses across metrics. These research findings will advance the intelligence and real-time capabilities of DD decision-making, and are expected to be extended to large-scale networks.

Acknowledgments

I extend my heartfelt gratitude to Prof. Yuan Liu and Assoc. Prof. Pengdeng Li of my school for their numerous insightful discussions and valuable feedback on this proposal.

References

Anwar, A. H.; Kamhoua, C.; and Leslie, N. 2020. A game-theoretic framework for dynamic cyber deception in internet of battlefield things. In *Proceedings of the 16th EAI International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services (MobiQuitous '19)*, 522–526.

Beltrán-López, P.; Pérez, M. G.; and Nespoli, P. 2025. Cyber Deception: Taxonomy, State of the Art, Frameworks, Trends, and Open Challenges. *IEEE Communications Surveys & Tutorials*, 1–1.

Das, B. C.; Amini, M. H.; and Wu, Y. 2025. Security and privacy challenges of large language models: A survey. *ACM Computing Surveys*, 57(6): 1–39.

Florea, R.; and Craus, M. 2022. A Game-Theoretic Approach for Network Security Using Honeypots. *Future Internet*, 14(12).

Guembe, B.; Azeta, A.; Misra, S.; Osamor, V. C.; Fernandez-Sanz, L.; and Pospelova, V. 2022. The Emerging Threat of Ai-driven Cyber Attacks: A Review. *Applied Artificial Intelligence*, 36(1): 2037254.

Hausken, K.; Welburn, J. W.; and Zhuang, J. 2024. A review of attacker–defender games and cyber security. *Games*, 15(4): 28.

Javadpour, A.; Ja'fari, F.; Taleb, T.; Shojafar, M.; and Benzaïd, C. 2024. A comprehensive survey on cyber deception techniques to improve honeypot performance. *Computers & Security*, 140: 103792.

Li, X.; Wang, S.; Yang, Q.; and Liu, Y. 2025. Defensive Strategies Against Sandwich Attacks: A Review. In *Proceedings of International Conference on Blockchain, Artificial Intelligence, and Trustworthy Systems(BlockSys)*.

Lohn, A. J. 2025. *Anticipating AI's Impact on the Cyber Offense-Defense Balance*. Center for Security and Emerging Technology.

Pawlick, J.; Colbert, E.; and Zhu, Q. 2019. A Game-theoretic Taxonomy and Survey of Defensive Deception for Cybersecurity and Privacy. *ACM Comput. Surv.*, 52(4).

Sun, Y.; Tian, Z.; Li, M.; Su, S.; Du, X.; and Guizani, M. 2021. Honeypot Identification in Softwarized Industrial Cyber–Physical Systems. *IEEE Transactions on Industrial Informatics*, 17(8): 5542–5551.

Zhu, M.; Anwar, A. H.; Wan, Z.; Cho, J.-H.; Kamhoua, C. A.; and Singh, M. P. 2021. A Survey of Defensive Deception: Approaches Using Game Theory and Machine Learning. *IEEE Communications Surveys & Tutorials*, 23(4): 2460–2493.