Research Statement

Yinuo Du
School of Computer Science
Carnegie Mellon University
yinuod@andrew.cmu.edu

Cybersecurity is a major concern with serious implications for society and the economy. Coming from a computer engineering background, I am fascinated by the complex problem of cybersecurity, which involves the strategic interaction among attackers, defenders, and end users. Both human factors and artificial intelligence are crucial to building effective cybersecurity systems; understanding human behavior and leveraging AI techniques can significantly enhance our protective measures. My curiosity about social decision science and artificial intelligence drives me to explore interdisciplinary research, combining insights from these domains to address cybersecurity challenges more effectively.

My research agenda focuses on **human and AI decision-making in cybersecurity**. My approach is fundamentally human-centered and adopts a multiagent perspective that unifies AI system development and human behavior research. I start by examining the fundamental principles of human learning and decision-making. When multiple humans and AI agents interact strategically, the complexity of these interactions increases. I use game theory to dissect and understand these strategic dynamics, drawing on behavioral insights to predict and influence outcomes. My goal is to leverage the strengths of both human and AI for cybersecurity, offering generalizable insights that can be applied in a wide range of situations.

My research themes are summarized below:

- **Game theory, Reinforcement Learning, and Cognitive Modeling for Cyber Defense.** A significant research problem in cybersecurity is predicting attacker actions for proactive and adaptive defense, especially given the lack of large datasets of human decisions in realistic cyberdefense situations. My research formulates the interaction as strategic games between attacker and defender and uses cognitive modeling and reinforcement learning to better understand the attacker and design defense strategies for the defender. Through interactive games, I collect empirical data on human attacker behavior [3]. By using cognitive modeling, which requires less data to characterize human decision patterns, I can effectively understand and predict attacker behavior. I developed cognitive models that model human decision-making, providing insights into attacker behavior [2, 7]. I also apply reinforcement learning algorithms to these games to develop adaptive defense strategies that predict attacker actions and optimize resource allocation [1].
- **Complementary Human-AI Teaming for Cybersecurity and Beyond.** Many professional sectors in incident response and cyber operations are seeking to increase the use of AI agents as a solution to current expertise and capability gaps. However, incorporating AI into these workflows faces challenges due to the dynamic nature of team operations and directives. Innovative approaches are needed to integrate AI agents into the workflow of professional teams that continuously receive tasks and restraints from external sources. I explore novel forms of human-AI partnerships in network defense and conduct human-subject experiments to measure the team effectiveness [6].
- **Group Dynamics in Cooperation and Consensus.** Understanding and facilitating teamwork within groups is a fundamental challenge across various domains, from tactical coordination to conflict resolution across organizations. In cybersecurity, effective collaboration among defenders through threat intelligence sharing and achieving strategic consensus is essential. My research investigates how group dynamics influence the effectiveness of these collaborative efforts [5, 4]. In addition, I also compares the collective decision-making processes of human and large language model (LLM)-agent groups [8] to identify key determinants of successful collaboration.

My work is driven by real-world problems in cybersecurity and requires insights from across disciplines, such as social and decision science, artificial intelligence, and human-computer interaction. To communicate the impact of my results back into the world, I have developed connections with cybersecurity researchers from the Peraton Labs, and the Army Research Lab, and I have been **invited to give talks** at INFORMS, the Future of Cyber Deception Workshop, Cylab Partners Conference, Women in Cybersecurity, and AI for Social Decision-Making Institute.

Current Work

As shown Table 1, my work employs environments reflecting various levels of functional fidelity regarding computer networks. The agents within cyber environments have varying degrees of human fidelity regarding attackers, defenders, and end users. The choice of environment and agent fidelity is made to address the specific goals and challenges of my research themes. Lower fidelity simulated environments allow for faster experimentation, data collection, and exploration of fundamental principles. As my research moves towards more realistic and complex cybersecurity scenarios, the environments and agents become more functionally and behaviorally advanced.

For example, in my work on game theory, reinforcement learning, and cognitive modeling for cyber defense (C1), the lower fidelity simulated environments and heuristic/cognitive agents enable me to efficiently test and refine adaptive defense strategies. On the other hand, the human-AI teaming research (C2) requires higher fidelity emulated environments and real human participants to faithfully capture the dynamics of teamwork in cybersecurity operations. By spanning this spectrum of functional and human fidelity, I can derive generalizable insights that bridge the gap between theoretical models and real-world cybersecurity challenges.

	Functional Fidelity of the Environment		
Human Fidelity	**Low** (Simulation)	**Medium** (Advanced Simulation)	**High** (Emulation)
Low (Heuristic and Cognitive agents)	† RL for Cyber deception [1], † Cognitive Modeling of attacker [7]		
Medium (Human participants with moderate expertise)	Group † Consensus [8] & ‡ Cooperation[4, 5]	Individual [3] & ‡/⋄ Team [6] Defense Game	
High (Expert hackers and Cyber analysts)		⋄ Adversary biases Computational modeling, & Exploitation	‡ Empirical Evaluation & ⋄ MARL for Adaptive defense

Table 1: †: finished work; ‡: ongoing work; ⋄: future work

C1: Game theory, Reinforcement Learning, and Cognitive Modeling for Cyber Defense. Cyber Deception techniques have been through a significant growth in the past decades. To make the best use of security resources, game theory has been used to decide the combinations of techniques to use and their respective parameter settings. However, most previous works, such as Stackelberg security games for honeypot deployment and network hardening, assume the decision-making to be one-shot and focus on the decisions made when the resources are first deployed. In reality, the cyber network's security status evolves at each time instant, due to exogenous random events and attacker's actions. It is essential to continuously reassess the defense measures.

We proposed **the first two-player Markov game model** that captures the progressive nature of attacks and allows the defender to adaptively allocate defense resources in real-time [1]. We studied the use of RL to solve the game. Simulation experiments suggest that the RL-based strategy outperforms expert-designed heuristic defense strategies against various types of attackers. Currently, I am working with masters and undergraduate students to build **high-fidelity testbeds** using real-world enterprise servers in cloud platforms. These testbeds enable us to deploy and **empirically evaluate** active defense strategies in realistic scenarios, where defenders can dynamically adjust network configurations and deceptive resources in response to ongoing attack campaigns.

To better understand the behavior of human attackers, I designed a **human-like attacker** [7] based on Instance-Based Learning Theory, which is dynamic, adaptable, and able to learn from experience. The cognitive attacker constructs memory instances using the concept of cyber kill chain and keeps track of the resources it occupies (e.g., detected, scanned, exploited, and impacted hosts). Unlike optimal attackers, this model incorporates human cognitive constraints and exhibits realistic biases observed in actual cyber attackers. Simulation experiments show that it forces defenders to choose from a larger option space thus hinders the learning speed of the opponent. We designed an online **individual defense game (IDG)** [3] and confirmed that cognitive attacker agents are more challenging than carefully crafted optimal attack strategies for the most efficient human defenders.

C2: Complementary Human-AI Teaming for Cybersecurity and Beyond. The cycle of cyber incident response includes several stages, each with distinct objectives and requiring different human-AI team configurations. These stages typically include the *Triage* phase, where potential incidents are identified and assessed, and the *Engagement* phase, where the incident is contained, threats are removed, and systems are restored. In the *Triage* stage, the task is primarily classification, with success measured by accuracy and efficiency. Here, human strengths in leveraging contextual information complement the AI's ability to process large datasets. In contrast, the *Engagement stage* involves sequential decision-making, aiming to maximize the probability of identifying and responding to adversarial actions. This stage benefits from the AI's computational speed and predict potential threats, while humans provide strategic oversight and adaptability. To explore potential human-AI team defense paradigms at the engagement stage, we designed a **team defense game (TDG)**. In this game, human participants are paired with autonomous agents to protect a fictitious computer network against external malicious activity. This setup allows us to study and optimize the collaboration between human and AI agents in a realistic and dynamic cyber defense scenario. After consulting with experts, we learned that a cyber protection team usually consists of several members with similar expertise to share the responsibility of network defense, and lower-rank officers need to report to their superiors for approval before taking risky actions. To reflect this we designed TDG, such that the human is a semi-supervisor of the autonomous agent, which has the agency to take low-risk actions but requires human approval for resource intensive actions. In our human-subject experiment, we compared teams with three types of autonomous agents. We found that (1) agents with high competency are rated to be more trustworthy, and (2) agents adapting to their human partners lead to better team performance.

C3: Group Dynamics in Cooperation and Consensus. The projects described above focus on the interaction between attackers and defenders. Additionally, I have studied the dynamics among groups of human defenders. Using a multiagent modeling approach, I begin with *environment simulation* and *individual decision-making modeling*. This method enables us to test a wide range of hypotheses regarding environmental conditions that promote specific behaviors and innate cognitive mechanisms.

With this approach, I explored the factors that might affect **information sharing** among a group of defenders in collaboration with Prof. Palvi Aggarwal (University of Texas at El Paso) [5, 4]. We designed an online game that simulates the scenario as an iterated group prisoner's dilemma game and conducted human-subject experiments with four conditions of different incentive structures and information levels. I then built cognitive models based on various hypotheses about social value orientation. By comparing synthetic and human data, we can better explain the emergence of group dynamics. Our findings can also inform the threat intelligence sharing protocols among enterprises in the real world.

I also studied the communication aspect in groups. In collaboration with Prof. Prashanth Rajivan (University of Washington), we studied the **group consensus** decision-making process [8]. I proposed a **free-form discussion algorithm** that allows multiple language agents to converse without external moderation. We then conducted **comparative conversation analysis** on the corpus of human and large language model-based agent groups. Our insights can help improve the collective problem-solving of cyber protection teams.

Future Plan

In the short term, I plan to expand my current research themes along complementary human-AI teaming and reinforcement learning for cyber deception toward higher human and functional fidelity. In the longer term, I plan to leverage my expertise in computational modeling to better understand and exploit adversary biases.

F1: Complementary Human-AI Team Defense: I will continue my work in human-AI teaming through 1) Improving and expanding the TDG platform for better alignment with the nature of task and environment of various cyber scenarios; 2) Developing team-aware, controllable, communicative autonomous agents with capabilities complementary to human defenders in the various cyber scenarios; I plan to leverage cognitive models to predict human decisions (including the reliance-related decisions based on the human mental model of AI). I will validate and adapt existing metrics of complementarity and trust calibration to assess the effectiveness of H-AI teams.

F2: Multiagent Reinforcement Learning for Adaptive Cyber Defense: I plan to advance my research in cyber deception towards scalable real-world deployment by developing sample-efficient (multi-agent) algorithms for deception. To bypass the gap between defense strategies trained in simulated environments and real-world network, I will explore the application of model-based MARL to enable sample efficient training. Finally, networks in the real world are attacked by multiple attackers with diverse motivations and different levels of attack capabilities. I will extend the game model to a multi-attacker multi-defender game and study the application of multi-agent RL for cooperative defense.

F3: Adversary biases, Computational modeling, and Exploitation: Attackers are subject to cognitive limitations and bias, which leads to time-consuming mistakes or non-optimal sequences of decisions, creates opportunities to frustrate, curtail, or halt the attack completely. The first step is to model and theoretically analyze the performance of adversaries who engage in exploitation activities that require sequential decision-making in the presence of uncertainty. Decision-making can be affected by multiple biases operating at the same time. I plan to investigate how do multiple biases interact to produce the overall outcome, and how certain natural behavioral phenomena arise when the attackers exhibit one or multiple biases. I will study the network configurations that present the worst case for the adversaries.

Given these points, I intend to build a strong research team consisting of several graduate students and I plan to get support from external research funding. During my PhD at the Dynamic Decision Making Laboratory and AI for Social Good Lab at Carnegie Mellon University, my research has been supported by research grants from the Army Research Office (ARO). I have gained experience contributing to white papers for FutureEnterprise@CyLab and IARPA. Through such experience, I have learned how to write grant proposals and how to target research problems that are of interest to funding agencies. I plan to pursue research funding for the National Science Foundation's CAREER awards under the Secure and Trustworthy Cyberspace (SaTC) division, the Cyber Training division, and the Army Research Office (ARO).

[1] **Yinuo Du**, Zimeng Song, Stephanie Milani, Cleotilde Gonzales, and Fei Fang. "Learning to play an adaptive cyber deception game". In: *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems. Auckland, New Zealand*. Vol. 6. 2022.

[2] Tyler Malloy, **Yinuo Du**, Fei Fang, and Cleotilde Gonzalez. "Accounting for Transfer of Learning Using Human Behavior Models". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 11. 1. 2023, pp. 115–126.

[3] Baptiste Prebot, **Yinuo Du**, and Cleotilde Gonzalez. "Learning about simulated adversaries from human defenders using interactive cyber-defense games". In: *Journal of Cybersecurity* 9.1 (2023).

[4] **Yinuo Du**, Palvi Aggarwal, Kuldeep Singh, Fei Fang, and Cleotilde Gonzalez. "A Model of Human Behavior in Group Prisoner's Dilemma Games". In: *Cognitive Science* (2024). under preparation.

[5] **Yinuo Du**, Palvi Aggarwal, Kuldeep Singh, Fei Fang, and Cleotilde Gonzalez. "Emergent Cooperative Behavior in Group Prisoner's Dilemma Games". In: *Decision* (2024). under preparation.

[6] **Yinuo Du**, Baptiste Prebot, Tyler Malloy, Fei Fang, and Cleotilde Gonzalez. "Experimental Evaluation of Cognitive Agents for Collaboration in Human-Autonomy Cyber Defense Teams". In: *Computers in Human Behavior: Artificial Intelligence* (2024). under review.

[7] **Yinuo Du**, Baptiste Prebot, Tyler Malloy, and Cleotilde Gonzalez. "A Cyber-War Between Bots: Cognitive Attackers are More Challenging for Defenders than Strategic Attackers". In: *ACM Transactions of Social Computing* (2024). Just Accepted.

[8] **Yinuo Du**, Prashanth Rajivan, and Cleotilde Gonzalez. "Large Language Models for Collective Problem-Solving: Insights into Group Consensus". In: *Proceedings of the Annual Meeting of the Cognitive Science Society, 46 (0)*. 2024.