Department of Data Science
香港城市大學
City University of Hong Kong

DS SEMINAR

Advancing Cybersecurity with Agentic, Scalable, and Reliable AI

Date: 26 January 2026 (Monday)

Time: 9:30am - 10:30am

Zoom: https://cityu.zoom.us/j/87219327253

Mr. Jiahao YU

ABSTRACT

Modern society is built on a foundation of complex software, yet our ability to secure it is fundamentally outpaced by the scale and speed of emerging threats. The current paradigm relies on human experts to manually discover, analyze, and patch vulnerabilities—a process that is increasingly unsustainable against automated attacks. In this seminar, I present a vision for the next generation of cybersecurity: autonomous AI systems capable of managing the entire vulnerability lifecycle, from discovery to remediation.

I will discuss my research framework, which addresses the core challenges of AI security through three interconnected thrusts: Agency, Scalability, and Reliability. First, I will demonstrate how to build Agentic systems that mimic the multi-step reasoning of human security experts. I will introduce PatchAgent and BandFuzz, autonomous frameworks that utilize reinforcement learning and tool use to reason through complex bugs, achieving state-of-the-art results in automated program repair and fuzzing. Second, I will address the "small data" challenge in security by introducing Scalable learning techniques. I will present GPO (Generative Preference Optimization) and EntroPO, algorithms that scale performance through synthetic data generation and test-time compute scaling, enabling smaller models to outperform proprietary giants on reasoning benchmarks. Finally, I will discuss Reliability, ensuring that these powerful agents are trustworthy. I will cover my work on LLM-Fuzzer, the first automated red-teaming framework for Large Language Models, and AIRS, a method for explaining the decisions of Deep Reinforcement Learning agents in security contexts.

This research has resulted in tangible real-world impact, including a semi-final win in the DARPA AI Cyber Challenge (AIxCC) securing a $3 million prize, and the adoption of my red-teaming tools by industry leaders such as Microsoft, OpenAI, and ByteDance. I will conclude by outlining a future roadmap for creating unified, interpretable, and collaborative AI agents that will define the future of digital defense.

GUEST SPEAKER'S PROFILE

Jiahao YU is a Ph.D. candidate in Computer Science at Northwestern University, advised by Professor Xinyu Xing. His research operates at the intersection of AI and Cybersecurity, with a focus on developing autonomous agents that are Agentic, Scalable, and Reliable. His work addresses critical challenges in automated vulnerability management, from red-teaming Large Language Models (LLMs) to autonomous patching and fuzzing.

Jiahao's research has produced widely adopted tools such as LLM-Fuzzer, which is currently used by internal red teams at Microsoft, OpenAI, Meta, and ByteDance. He was a core member of the team that won the semi-final of the DARPA AI Cyber Challenge (AIxCC), securing a $3 million prize and discovering the most zero-day vulnerabilities in the final competition. His work has been published in top-tier conferences including NeurIPS, ICML, ICLR, and USENIX Security. Prior to Northwestern, he received his B.S. from Shanghai Jiao Tong University.

Enquiries: ds.go@cityu.edu.hk

All are welcome