Unleashing the Power of AI to Automate Cybersecurity

Luca Gioacchini

The exponential growth of interconnected devices and networks has created unprecedented cybersecurity challenges, with an impressive amount of estimated Internet users and networked devices in the last decade that are victims of cyberattacks. While driving innovation, this hyper-connected ecosystem has simultaneously expanded the attack surface for cyber threats. As cyber attackers leverage Artificial Intelligence (AI) to perform aggressive campaigns at scale, traditional, mostly manual approaches to cybersecurity have become increasingly inadequate. This thesis investigates AI-based solutions for two critical cybersecurity domains: Network Traffic Analysis (NTA) and Penetration Testing.

Concerning NTA, I address the challenge of detecting coordinated attacks through two main contributions. First, I introduce i-DarkVec, a novel methodology that applies Natural Language Processing techniques to generate self-supervised representations of the traffic sent by network hosts. This approach demonstrates exceptional accuracy in categorising coordinated activities while achieving remarkable computational efficiency compared to existing solutions, reducing training time from days to seconds. Through an unsupervised analysis, i-DarkVec successfully uncovered numerous previously unidentified coordinated activity groups, proving its value as a network monitoring tool.

Second, I explore the generalisation of the produced network traffic representations across different environments through two complementary approaches: Transfer Learning and Model Stacking. I develop two solutions to transfer the knowledge across networks within a customer/provider paradigm: an explicit alignment approach utilising common data points between environments, and a canonical transfer approach enabling knowledge transfer between heterogeneous networks. Furthermore, I demonstrate the effectiveness of model stacking in combining diverse embeddings that capture complementary aspects of network traffic, from connectivity patterns to temporal dynamics and information about traffic volume and type. This comprehensive representation leads to significant performance improvements in cross-network scenarios.

Concerning penetration testing, I investigate the potential of generative AI agents to automate security assessments. I present AgentQuest, a modular benchmarking framework that introduces novel behavioural metrics for evaluating generative agents progress toward the solution of reasoning-demanding tasks. Building upon this foundation, I develop AutoPenBench, a comprehensive benchmark for assessing autonomous penetration testing capabilities of generative AI agents. The evaluation of both a fully autonomous agent and a semi-autonomous agent reveals current limitations and future possibilities in AI-driven security testing.

While fully autonomous agents shows limited effectiveness, particularly in complex real-world scenarios, semi-autonomous agents operating under human supervision demonstrate remarkable improvements. These findings suggest that while complete automation remains a distant goal, AI can effectively assist human analysts and augment their expertise in cybersecurity tasks. The semi-autonomous approach significantly reduces the workload of security professionals while maintaining high-quality results, establishing a practical path forward for AI-assisted cybersecurity.

This research makes significant contributions to both theoretical frameworks and practical tools for automated cybersecurity. The findings demonstrate the potential of AI to enhance threat detection and security testing while acknowledging the continued importance of human oversight in complex security assessments. The developed methodologies and tools provide a foundation for future research in AI-driven cybersecurity, offering scalable solutions to address the evolving challenges of modern cyber threats.