COMPUTER AND CONTROL ENGINEERING

DAUIN - Agentic AI for Cybersecurity: Autonomous Red-Blue Agents in Interactive Cyber Environments

Funded By	Dipartimento DAUIN
Supervisor	MELLIA MARCO - marco.mellia@polito.it
Contact	MELLIA MARCO - marco.mellia@polito.it VASSIO LUCA - luca.vassio@polito.it
Context of the research activity	The project explores whether autonomous AI agents can replicate and strengthen the traditional Red, Blue, and Purple cybersecurity cycle. Red agents will generate realistic attack scenarios, while Blue agents analyse incidents and produce mitigations. A central Purple agent will observe interactions, extract knowledge, and iteratively improve both sides through coaching or learning mechanisms. By creating a unified environment where agents evolve together, the project investigates how far this competitive loop can be pushed, including the possibility of discovering novel attacks and defence strategies, minimising human intervention.
	Modern cybersecurity relies on the dynamic interplay between Red teams (attackers) and Blue teams (defenders). Today, these roles are almost entirely human-operated, limiting scalability and slowing the response to rapidly evolving threats. Meanwhile, LLMs and autonomous AI agents are showing strong performance in reasoning, code generation, and cybersecurity tasks. This creates a unique opportunity: can we automate the full attack–defence cycle, enabling continuous adversarial testing and rapid adaptation? Current research on Red or Blue agents is fragmented, often manually crafted, and evaluated only in simplified settings. No unified framework exists where Red and Blue co-evolve under structured oversight. To automate this loop, we introduce the purple agent, with coaching and coordination goals to guide the Red and Blue agents in improving their abilities. Automating this loop could accelerate vulnerability discovery, improve defensive robustness, and support safer, data-driven security evaluation for real-world systems. This project aims to explore the development of an ecosystem of autonomous agents capable of challenging, analysing, and improving each other. The candidate will investigate whether recent advances in large language models (LLMs) and autonomous agents can reproduce and enhance the classical Red–Blue cycle used in cybersecurity operations.

Objectives	The objectives are: 1. Design autonomous Red and Blue agents capable of generating realistic attacks and analysing incidents, respectively, moving beyond current prompt-engineered, isolated prototypes. 2. Develop a supervising Purple agent that collects execution traces, audits interactions, and improves both sides through coaching, reinforcement learning, episodic memory, or architectural modifications 3. Create a unified environment enabling continuous adversarial interaction where Red challenges Blue and Blue adapts, forming a self-reinforcing learning loop. 4. Assess emergent capabilities, including the popotentialor agents to discover novel attack vectors or defence strategies not explicitly coded by humans. 5. Quantify improvements over baseline agents and evaluate whether autonomous co-evolution can lead to measurable gains in mitigation quality, reasoning clarity, and knowledge accumulation. Research work plan: We foresee three phases: • During the first year, the candidate will review the state of the art, and focus on the initial design of the red and blue agents in isolation, with the definition of benchmarks to properly assess their performance. • During the second year, the candidate will focus on the setup and design of the “coaching loop” in which Purple analyses the Red and Blue reasoning traces and proposes improvements. This will ininvolvehe study of mechanisms for adaptive attack generation for Purple’s evaluation. • During the third year, the candidate will deep dive into AgenticAI approaches, fine-tuning the agents with richer memory, external memory, and improved architectures while investigating RL-based or curriculum-learning schemes that gradually increase attack sophistication. References: - F. De Santis, K. Huang, K., R.Valentim, D. Giordano, M. Mellia, B. Houidi, D.Rossi, CFA-bench: Cybersecurity Forensic LLM Agent Benchmark and Testing. Proceedings of the 10th International Workshop on Traffic MMeasurementsfor Cybersecurity, 2025. - Gioacchini, L., Delsanto, A., Drago, I., Mellia, M., Siracusano, G., & Bifulco, R. (2025, novembre). AutoPenBench: A Vulnerability Testing Benchmark for Generative Agents. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: IndustryTrack (pp. 1615-1624). - Fumero, S., Huang, K., Boffa, M., Giordano, D., Mellia, M., Houidi, Z. B., & Rossi, D. (2025). CyberSleuth: Autonomous Blue-Team LLM Agent for Web Attack Forensics. arXiv preprint arXiv:2508.20643.

Skills and competencies for the development of the activity	• Strong programming skills in Python, with experience in AI frameworks (PyTorch, TensorFlow) • Understanding of vulnerabilities, and common attack and defence techniques • Interest or experience in AI agent design, LLM fine-tuning • Familiarity with tools such as Metasploit, Wireshark, Snort, or Cuckoo Sandbox is a plus