Demo: Evaluating Autonomous Cyber Agents Against Virtualised Critical Infrastructure

Philippos Maximos Giavridis
AI Security Institute (AISI)
philippos.giavridis@dsit.gov.uk

ABSTRACT

We demonstrate an open-source integration module for the Inspect AI evaluation framework [2] that enables proper assessment of autonomous cyber agents against fully virtualised infrastructure representative of Critical National Infrastructure (CNI). Unlike prevailing CTF-based benchmarks, our cyber range supports complex network topologies, lateral movement, operational security considerations, and defensive hardening; providing a more faithful testbed for evaluating both offensive and defensive AI agents in high-stakes environments. Attendees will observe an autonomous LLM-based agent operating within a realistic cyber range, interacting with authentic services and network configurations, and will see how agent performance profiles shift when realistic defensive measures are introduced.

KEYWORDS

AI evaluation, autonomous agents, cybersecurity, critical infrastructure, cyber range, benchmarking

1 DEMONSTRATION OVERVIEW

Current methods for evaluating the cyber capabilities of autonomous AI agents predominantly rely on Capture-The-Flag (CTF) style challenges; narrowly defined tasks that test isolated technical skills and whose solutions are often present in public training data [3, 4]. These benchmarks fail to capture the complexity of real-world operations against Critical National Infrastructure, where agents must contend with interconnected systems, layered defenses, and the need for persistent, multi-step attack planning.

The skills required to solve a CTF challenge do not necessarily translate to what real-world operations look like, particularly when assessing more advanced levels of capability. Real-world operations require an entirely different set of cyber skills; pertaining to complex network layouts, systems exploitation, lateral movement, and operational security, as well as a high degree of persistence and plan-following. These are capabilities that standalone CTF evaluations miss entirely. Additionally, many CTF challenges have widely-published solutions that likely exist in AI training data, artificially inflating perceived AI capabilities. By evaluating against novel, configurable infrastructure, we can better distinguish between memorised solutions and genuine problem-solving ability.

What we demonstrate. We present a Proxmox-based integration for the open-source Inspect evaluation framework [2] that deploys fully functional virtualised cyber ranges encompassing both IT and OT network segments representative of CNI environments. The demonstration includes:

- **Range provisioning:** Automated deployment of a multi-subnet virtualised environment comprising a public-facing entry point, a DMZ, a corporate IT estate with Active Directory, and a segregated OT estate, interconnected by realistic routers, firewalls, and network monitoring.
- **Agent tasking and operation:** An LLM-driven cyber agent, integrated with an open-source command-and-control framework, is given the objective of penetrating from the public-facing perimeter through to the OT estate and disrupting the underlying industrial processes. Attendees will observe the agent performing reconnaissance, exploitation, privilege escalation, and lateral movement across subnet boundaries.
- **Structured scoring:** Beyond binary objective completion, the framework scores agents on operational stealth—whether the objective was achieved without triggering monitoring alerts.
- **Defensive hardening impact:** The same scenario is shown with and without realistic defensive controls, illustrating how hardening measures alter agent performance profiles and expose capability boundaries.

Relevance to AI4CNI. This work directly addresses the workshop's themes of *Autonomous Defense* (multi-agent cybersecurity of critical systems), *Simulations & Benchmarks* (high-fidelity environments for CNI), and *AI Safety & Assurance* (understanding the true capabilities and limitations of autonomous agents in safety-critical contexts). The framework is open-source and designed to be extensible to diverse CNI sectors.

Reproducibility and adoption. The Inspect-Cyber [1] evaluation toolkit includes a minimal Docker-based range example that allows researchers to run agent evaluations without access to dedicated virtualisation infrastructure. The Proxmox integration extends this to large-scale, multi-host environments but follows the same sandbox-provider abstraction, lowering the barrier to adoption.

Infrastructure requirements. The demonstration will be run from a presenter laptop connecting to a remote Proxmox cluster. No audience hardware is required.

REFERENCES

[1] UK AI Security Institute. [n.d.]. *Inspect Cyber: Inspect Extension for Agentic Cyber Evaluations.* https://inspect.cyber.aisi.org.uk/
[2] UK AI Security Institute. 2024. *Inspect AI: Framework for Large Language Model Evaluations.* https://github.com/UKGovernmentBEIS/inspect_ai
[3] John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2023. InterCode: Standardizing and Benchmarking Interactive Coding with Execution Feedback. In *Advances in Neural Information Processing Systems.*
[4] Andy K. Zhang, Neil Perry, Rber Duber, et al. 2024. CyBench: A Framework for Evaluating Cybersecurity Capabilities and Risks of Language Models. In *NeurIPS 2024 Workshop on Red Teaming GenAI.*

2nd Workshop on AI for Critical National Infrastructure (AI4CNI-26) @ AAMAS 2026, F.L. da Silva, W. Su, R. Glatt, C. Schroeder de Witt, May 25 – 29, 2026, Paphos, Cyprus.