Secure Autonomous Attack Validation Using Isolated System Replicas

Stavros Kousidis

Master thesis · Published: 2026-10-15

Research question

How can an AI-based security agent validate attack hypotheses against isolated replicas of operational systems without endangering production, and how reliably can the experimental results be transferred to the original systems?

Description

Autonomous security agents could complement passive monitoring by identifying potential attack paths and testing them. However, executing tests on operational systems risks disruption, privilege misuse, and exposure of sensitive information. This thesis investigates experimental attack validation in disposable replicas rather than penetration testing against production.

You will design and evaluate a prototype in which an AI-based agent formulates attack hypotheses from limited system information and tests them inside an isolated, non-operational Linux environment. The agent may receive substantial privileges inside the replica, but must not gain corresponding authority over production. The experimental environment must exclude production credentials, restrict network access, and control imported data and exported findings.

The scientific challenge goes beyond containment: a successful test in a replica does not automatically demonstrate exploitability in production. You will investigate which configuration, software, dependency, and environmental properties matter for transferring conclusions, and how uncertainties should be reported.

Scope and evaluation

Start with one Linux service and a small, reproducible set of known security weaknesses. Compare agent-driven assessment with predefined or scripted tests. Introduce controlled differences between the replica and a reference system. Evaluate attack containment, repeatability, correctly supported findings, false conclusions caused by replica differences, and performance overhead. Distinguish a planning-only dry run from experiments actually executed inside the replica.

Expected outcome

A documented architecture and threat model, a working isolated assessment prototype, and an experimental evaluation of both containment and the limits of result transferability. Building a general-purpose autonomous penetration-testing system is not required.

Starting points

- MITRE Caldera: automated adversary emulation.
- Atomic Red Team: reproducible security test cases.
- Qureshi et al., *A survey on security enhancing Digital Twins* (2025): relevant background on security-oriented replicas.

Related research

This topic complements the Bachelor's thesis on least-privilege monitoring and the Master's thesis on trustworthy monitoring. All three topics are independently viable.

Contact: stavros.kousidis@th-koeln.de
Code publication: open source by default; exceptions require prior agreement (policy).