Design and Implementation of a Least-Privilege Security Monitoring Agent for Linux

Stavros Kousidis

Bachelor thesis · Published: 2026-10-15

Research question

How can an AI-based security monitoring agent be isolated from a monitored Linux system while retaining sufficient information to detect and analyze security-relevant events?

Description

AI-based security agents can support intrusion detection and incident analysis but introduce additional security risks, particularly when granted privileged access to the monitored system. This thesis investigates an architecture that separates event collection, analysis, and reporting according to the principle of least privilege.

You will design and implement a prototype consisting of a minimal Linux event collector, an isolated AI-based analysis component, and a controlled reporting interface. The analysis agent operates without privileged access to the host and communicates only through explicitly defined interfaces.

The security evaluation will examine sandbox isolation, resource restrictions, and attacks involving maliciously crafted log entries, including prompt injection. A rule-based analyzer provides a baseline for comparison. The focus is on the security architecture, not on training an AI model.

Expected outcome

A working Linux prototype, a documented security architecture, and an experimental evaluation of isolation guarantees and limitations.

Starting points

- Shah, *LogJack: Indirect Prompt Injection Through Cloud Logs Against LLM Debugging Agents* (2026).
- OWASP, *LLM Prompt Injection Prevention Cheat Sheet.*
- Linux seccomp documentation.

Related research

This is part of Secure Monitoring with Untrusted AI Agents. A related Master's thesis investigates the trustworthiness of monitoring when individual components may be compromised. Both theses can be pursued independently.

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