Trustworthy Security Monitoring with Isolated and Potentially Compromised AI Agents

Stavros Kousidis

Master thesis · Published: 2026-10-15

Research question

How can the integrity and availability of security monitoring be maintained or failures reliably detected when attackers can manipulate monitoring data, compromise analysis agents, or interfere with the monitoring infrastructure?

Description

Security monitoring systems must remain trustworthy even when individual components are compromised. While sandboxing can prevent an AI-based analysis agent from directly affecting the monitored system, it does not guarantee the integrity of monitoring data, the correctness of analysis results, or the availability of monitoring infrastructure.

This thesis investigates security architectures that separate event collection, evidence verification, AI-based analysis, and reporting into distinct trust domains. Building on a minimal prototype, you will investigate independent verification mechanisms, potentially involving an external monitoring component, authenticated event streams, sequence validation, and monitoring-failure detection.

A systematic threat model will distinguish attackers who manipulate log entries, compromise the analysis agent, or obtain privileged access to the monitored host. Different architectures will be evaluated experimentally for attack containment, evidence integrity, monitoring-failure detectability, and performance overhead.

Particular attention will be given to which security guarantees are achievable under each attacker model and to the fundamental limits of monitoring a potentially compromised host.

Expected outcome

A precise threat model, a prototype with independent monitoring and verification components, and a comparative security evaluation identifying achievable guarantees and their limitations.

Starting points

- Schneier and Kelsey, *Secure Audit Logs to Support Computer Forensics* (1999).
- NIST SP 800-92, *Guide to Computer Security Log Management.*
- Shah, *LogJack: Indirect Prompt Injection Through Cloud Logs* (2026).

Related research

This is part of Secure Monitoring with Untrusted AI Agents. A related Bachelor's thesis investigates how to isolate an AI monitoring agent from its host. The Master's thesis is independently viable and does not depend on that prototype being completed.

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