AI AGENTS FOR DATA LEAK PREVENTION: A FRAMEWORK FOR CRITICAL INFRASTRUCTURE PROTECTION

Miloslava Plachkinova
Kennesaw State University
mplachki@kennesaw.edu

Ace Vo
Loyola Marymount University
ace.vo@lmu.edu

ABSTRACT

Artificial intelligence (AI) agents are transforming critical national infrastructure by enabling automation, predictive analytics, and real-time decision-making across sectors such as healthcare, energy, finance, and transportation. However, these capabilities introduce new vulnerabilities, particularly in data leak prevention (DLP). As AI agents process and interact with sensitive data, traditional DLP mechanisms struggle to address the scale, complexity, and dynamic nature of these environments. In this paper, we develop and evaluate a novel DLP framework tailored to AI agents in healthcare using a design science approach. We propose AI Sentinel Theory – a theoretical framework grounded in Socio-Technical Systems Theory. We incorporate an independent, dedicated AI system as a continuous monitor, auditor, and governor of operational AI behavior. Guided by sentinel principles (independence of oversight, context-aware detection, adaptive learning, and explainable enforcement), our framework provides proactive, real-time protection. We utilized qualitative data and case studies from the healthcare sector to evaluate our artifact and demonstrate its utility and effectiveness. Our findings highlight the need for hybrid security models and intelligent oversight to secure AI agent deployments in healthcare. We offer strategic recommendations for infrastructure operators, policymakers, and technology developers to balance innovation with resilience.

Keywords: Artificial Intelligence (AI), AI Agents, AI Sentinel Theory, Socio-Technical Systems Theory, Agency Theory, Critical National Infrastructure (CNI), Healthcare, Data Leak Prevention (DLP), Cybersecurity, Design Science Research (DSR)