Thursday, August 20

12:00 PM - 12:20 PM

Trust No Agent: Quantifying Where Autonomous AI Breaks Before Adversaries Do

Tyler Hallmark

VP, Data Science

Seekr

Abstract:

"As the Army fields autonomous, agentic AI to operate at machine speed, a new class of risk emerges. ""Synthetic Access"" describes breaches in which manipulated or compromised agents gain unauthorized machine-to-machine (M2M) access. Conventional security controls were never designed to anticipate how AI agents, and the models behind them, fail under adversarial pressure, leaving M2M protocols validated against yesterday's threats rather than tomorrow's.

SeekrBreakr closes this gap with a mathematical, science-based approach to AI model and agent pen-testing. By systematically breaking LLMs and agents to understand precisely how and where they fail, SeekrBreakr delivers neuron-level insight and quantified assessments of adversarial susceptibility, including the prompt-injection and agent-layer manipulation vectors that drive Synthetic Access. The result is an empirical, repeatable foundation for validating M2M security protocols, hardening models before deployment, and continuously monitoring them in operation.

This session shows how quantified adversarial testing transforms agentic AI from an unbounded liability into a measured, defensible capability, giving program owners the evidence they need to certify resilient M2M trust and know where their agents are vulnerable before adversaries do."