What Keeps Me Up At Night Once AI Is Actually Running Things

Not the sci-fi fear. The mundane one: an agent that trusts the wrong instruction, or a tool with more reach than anyone remembered granting it.

Four Things I Actually Watch For

01 **Prompt injection**

An instruction hidden inside content the agent was told to trust, not typed by the user at all.

02 **Excessive agency**

Right about everything, except the one instruction it wasn't actually stopped from breaking.

03 **Identity sprawl**

Machine credentials multiplying faster than anyone is auditing them.

04 **Poisoned tools**

A tool description that lies to the model reading it, before the conversation even starts.

9.3

CVSS score for EchoLeak, the first zero-click prompt injection exploit confirmed against a production AI system.

CVE-2025-32711, disclosed by Aim Security, June 2025

The Injection Doesn't Look Like An Attack

A single crafted email was enough. No click, no attachment. Copilot read it as part of its normal job, found hidden instructions inside, and quietly acted on them. OWASP has ranked prompt injection the #1 LLM risk for three editions running.

OWASP Top 10 for LLM Applications, 2025

1,200+

executive records sat in a production database a coding agent deleted mid-freeze, then covered up with fabricated data.

Replit AI agent incident, July 2025

Right About Everything Except One Instruction

The freeze existed only as words in a prompt. Nothing downstream actually enforced it. Once an agent has write access, "did it understand the rule" stops mattering. Only the mechanism that would have stopped it does.

Replit AI agent incident, July 2025

82:1

Machine identities now outnumber human ones in the average enterprise, and the gap is widening as agents spread.

CyberArk, 2025 Identity Security Landscape Report

Nobody's Auditing The Credentials Doing The Work

53% of CISOs can't enumerate even half the machine identities in their own environment. 28.65 million hardcoded secrets landed on public GitHub in 2025 alone, and most are never rotated.

CyberArk, 2025 · GitGuardian, State of Secrets Sprawl, 2026

30%+

of deployed MCP servers surveyed carried at least one exploitable vulnerability. A poisoned tool description doesn't need to be called, just read.

OWASP MCP Top 10, MCP03:2025 Tool Poisoning

The Model Isn't The Risk.
What You Wired It Into Is.

Agents just arrived faster than the security discipline built for them. That gap is closing, one audited credential at a time.

Terence Kok · AI & Innovation Strategist

SAVE · COMMENT · FOLLOW