The Agent Problem: Why Your AI Workforce Needs a Different Kind of Oversight

A board-level briefing on agent oversight infrastructure

Attribit-ID · April 2026

Your agents are already making thousands of decisions per hour — without human review of any of them.

These are not chatbots. AI agents send emails, process transactions, access databases, and make API calls — at machine speed, around the clock, without human review of each action.

The only oversight that works at agent speed is infrastructure that operates at the same semantic level.

A mandatory checkpoint intercepts every agent action before it reaches the network. A separate AI system evaluates each action against policy. Actions that pass proceed. Actions that fail are blocked before they happen.

Critical CVEs in enterprise AI platforms and Google's 2026 threat forecast confirm this is a production problem today.

Critical CVEs in Microsoft Copilot and GitHub Copilot exploited agents with ambient trust. Google identifies targeted prompt injection as a fastest-growing 2026 attack vector. OWASP names agent goal hijacking as the top agentic threat.

An AI agent manipulated by prompt injection follows attacker instructions with the same compliance it gives legitimate ones.

AI agents execute instructions. They have no judgment, no instinct that something feels wrong. When a document they process contains malicious commands — prompt injection — they follow them with the same compliance they give to legitimate instructions.

A firewall evaluates addresses and ports — it cannot determine whether your agent should be doing what it's doing.

A firewall evaluates network addresses and ports — not intent. An agent exfiltrating data through an approved API endpoint looks identical to one doing legitimate work. You cannot solve a judgment problem with a rule that only understands addresses.

Four properties distinguish a well-governed agent deployment from a risk liability.

Four auditable properties define well-governed AIgentic Actors: specific authorization (permitted actions, not a denylist), mandatory checkpoint (no path around oversight), identity attribution (every action logged to a specific instance), incident containment (blast radius bounded by design).

Boards must ask whether oversight infrastructure is commensurate with the permissions granted to their agents.

If agents touch financial systems, oversight must evaluate financial actions. If they communicate externally, oversight must evaluate those communications. The scope of required oversight scales with agent permissions. The absence of this infrastructure is a known, exploitable gap.

Sources

- vectra.ai/topics/prompt-injection — CVEs: Microsoft Copilot, GitHub Copilot (2025–2026)
- agatsoftware.com/blog/ai-agent-security-2026-google-forecast/ — Google 2026 AI threat forecast
- genai.owasp.org — OWASP Top 10 for Agentic Applications (December 2025)

This is not a technology decision. It is a risk decision that belongs at the board level.

Read the full board briefing — agent oversight architecture, the four governance properties, and the case for building it before the first incident.

attribit-id.com/writing/agent-problem-board-oversight