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

The governance question your current risk framework wasn't built to answer

AI agents take real actions on your behalf — sending emails, processing transactions, accessing databases — at machine speed, without per-action human review. Traditional security tools cannot evaluate whether an agent should be doing what it's doing. This briefing explains the oversight architecture boards need to ask about.

What Makes Agents Different

An AI agent executes instructions. When those instructions are manipulated — by an adversary embedding malicious commands in a document the agent processes — the agent follows them with the same compliance it brings to legitimate ones. It has no instinct that something feels wrong.

Google's security researchers identify this as one of the fastest-growing attack vectors of 2026. Microsoft Copilot and GitHub Copilot have had critical vulnerabilities disclosed in the past year exploiting exactly this weakness. This is not theoretical.

The Architectural Response

The answer is not a better firewall. A firewall knows about network addresses and ports. It cannot evaluate whether your agent *should* be doing what it's currently doing.

The answer is oversight infrastructure operating at the same semantic level as the agents — a mandatory checkpoint every agent action passes through before it reaches the network. A separate AI system evaluates whether the action is consistent with defined policy. The agent cannot perceive this oversight layer, and therefore cannot reason around it.

Three Properties That Make It Work

Mandatory routing. Agent traffic is physically separated from human network traffic and can only reach external systems through the oversight checkpoint. This is a topological property — it holds even if the agent's software is compromised.

Specific authorization. Each agent is authorized to do specific things, and only those things. Not a list of prohibitions — a short list of permissions. Everything else is blocked.

Identity and accountability. Every action is logged against the specific agent instance that took it. Attribution reaches the instance, the task, and the moment — not just "the AI system did this."

The Board's Governance Question

The Board's Governance Question

The question is not "are we using AI agents?" It is: do we have oversight infrastructure commensurate with the permissions we've granted them?

Agents with access to financial systems need oversight capable of evaluating financial actions against policy. Agents with access to customer data need oversight capable of evaluating data handling. The scope of required oversight scales with the scope of agent permissions.

The absence of this infrastructure is a known, documented, exploitable gap — identified as the top vulnerability class in production agentic systems by more than 100 security researchers in December 2025.

Four Properties of a Well-Governed Deployment

Boards can ask about each of these and expect a specific answer:

- **Specific authorization** — each agent role has defined permitted actions, enforced technically
- **Mandatory checkpoint** — agent traffic cannot reach external systems without passing policy review
- **Identity attribution** — every action logged to the specific instance, task, and moment
- **Incident containment** — a compromised agent can be revoked immediately, without disrupting others

If your organization cannot answer all four affirmatively, risk exposure scales with agent permissions.

Two questions this answers

- What makes AI agents fundamentally different from the human workforce your governance framework was built for?
- What four properties define a well-governed agent deployment — and what should boards ask to verify them?

This is an engineering problem, not a philosophical impasse.

The oversight infrastructure exists. The architecture is understood. The question is whether your organization builds it before — or after — the first incident.

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