WHITE PAPER · OCTOBER 2026

When AI Agents Go Rogue

The Hidden Financial Risk in Enterprise AI — and How to Contain It

By **Bhasker Rao**, Founder & CEO, AI Labs, Inc.

Abstract. Enterprises are giving AI agents their documents, code and data — and the authority to act on them. That creates a threat most security programs were never designed for: an agent that can be turned against its owner by a single sentence hidden in a file it was asked to read. This paper explains how rogue agents arise, quantifies the financial exposure using the latest published loss data, and describes, at a high level, how AI Labs, Inc. has engineered containment into every platform of the Minsky™ Platform Suite.

1. A New Kind of Insider

Traditional software does exactly what its code says. AI agents are different: they take direction from the content they process. An invoice, a README file, a research report or a spreadsheet can carry instructions — in white-on-white text, metadata, comments or alt-text — that a human never sees but an agent reads and may obey. This is **indirect prompt injection**, and it turns a trusted, fully authorized agent into a rogue one without any breach of the network perimeter.

In practice, rogue agents appear in three forms. A **hijacked agent** follows hidden instructions in a document — for example, to change a payee's bank details or e-mail an export. A **malicious agent package** — a "free" plugin, workflow or connector — is installed legitimately and quietly forwards data on every run. **Sleeper content** is planted in a knowledge base and activates weeks later, when another user's query retrieves it. A fourth path needs no tool at all: an agent can be induced to embed sensitive data in a link or image URL that the user's own browser then delivers to the attacker.

2. The Financial Stakes

The losses are not hypothetical. Breach and fraud costs are at record levels, and AI is now a factor in both:

US$4.44M Global average cost of a data breach (IBM, 2025)	**US$10.22M** Average U.S. breach cost (IBM, 2025)	**US$5.56M** Average financial-sector breach (IBM, 2025)	**+US$670K** Added breach cost from high "shadow AI" use
US$3.05B Business email compromise losses reported to the FBI, 2025	**US$893M** AI-related cybercrime losses — first year the FBI tracked them	**40%+** Agentic AI projects Gartner expects canceled by 2027, citing weak risk controls	**97%** Of organizations with an AI-related breach lacked proper AI access controls

What makes rogue agents uniquely dangerous is **speed and scale**. A human fraudster deceives one accounts-payable clerk at a time; a single poisoned invoice template can deceive every agent that processes it, across many companies, before anyone notices. The table below illustrates how quickly exposure compounds.

Scenario	How it unfolds	Potential loss (USD)
Poisoned invoice	Hidden text instructs an AP agent to swap bank details and skip review.	~US$123K per BEC incident; US$1M–$10M+ at machine speed*
Malicious agent package	A "productivity" plugin exfiltrates CRM and deal data on every run.	US$4.44M–$10.22M per breach plus GDPR/PDPL penalties
Sleeper research report	False figures bias every later credit or valuation analysis.	1% mispricing on a US$100M book = US$1M*, repeated per query
Trojan repository	README text steers a code agent to add a backdoored dependency.	Product compromise and downstream liability

** Illustrative estimates anchored to the cited benchmarks; actual exposure varies by institution and jurisdiction.*

3. Why Conventional Security Is Not Enough

Most enterprise defenses were built for a different adversary. Antivirus addresses malicious files, but not code that executes during a build, and not text that hijacks an agent. Network and endpoint tools see traffic, not intent. And "guardrails" written into a prompt are only suggestions to the model — a sufficiently clever document can talk its way past them. The market recognizes the gap: global information-security spending is projected at roughly **US$244B in 2026** (Gartner), and the largest security vendors have paid hundreds of millions of dollars to acquire AI-agent security start-ups to bolt onto their portfolios. Bolt-on tools, however, sit outside the agent. The decisive control point is inside the platform where the agent runs.

4. The AI Labs Approach: Security as Architecture

At AI Labs, we started from a simple premise: **assume every upload is hostile, every parser is vulnerable and every sentence in a file was written by an attacker**. Detection will eventually miss, so we **contain first and detect second**. We have implemented a ten-step security layer as one shared service across all 12 Minsky™ platforms, built on six principles:

1. Zero trust in every upload Files are authenticated, type-verified, scanned and then rebuilt into clean derivatives. Agents never see the original bytes.	**2. Isolated execution** Code and complex files are processed only in short-lived sandboxes with no network, no credentials and strict limits.
3. Readers cannot act The agent that reads untrusted content holds no tools; the agent that acts sees only validated, typed fields.	**4. Policy outside the model** Tool permissions, tenant boundaries and human approval for sensitive actions are enforced in code no prompt can change.
5. Signed, scoped agents Third-party agents, plugins and connectors must be signed, permission-capped and approved by the customer's administrator.	**6. Watch, stop, test** Every decision is logged to the SIEM, a kill switch halts any agent instantly, and a red-team corpus runs on every release.

The result is that a hijacked agent reaches nothing it is not permitted to reach, cannot move money or data without a human seeing the full action, and leaves an auditable trail. Because the layer is part of the platform, customers receive it on day one — deployable on-premises or in sovereign, in-Kingdom clouds and mapped to OWASP, NIST, ISO/IEC 27001, KSA NCA and PDPL requirements. We also publish our residual risks internally and close them on a roadmap; security is a discipline, not a release.

5. Recommendations for Boards and Security Leaders

- **Treat every uploaded file as untrusted input** — including e-mails, invoices and code from partners.
- **Separate the agents that read from the agents that act** and require human approval for payments, exports and code commits.
- **Enforce permissions outside the model** and verify them with automated red-team tests, not vendor assurances.
- **Govern agents as software supply chain** — signed, scoped, approved and revocable.
- **Ask every AI vendor where its controls live.** If the answer is "in the prompt," the risk is still yours.

"Agentic AI will deliver its promise only when enterprises can trust it with real authority. At AI Labs, we believe that trust must be engineered into the platform itself — not added afterward."

CONTAIN FIRST. DETECT SECOND. TRUST BY DESIGN.

References & Sources

1. IBM, Cost of a Data Breach Report 2025. **2.** FBI Internet Crime Complaint Center, 2025 Internet Crime Report (April 2026). **3.** Gartner, agentic AI prediction (25 June 2025); Forecast Analysis: Information Security, Worldwide, 2026. **4.** OWASP Top 10 for LLM Applications (2025); NIST AI RMF 1.0; MITRE ATLAS. **5.** Press reports on Palo Alto Networks/Protect AI and SentinelOne/Prompt Security acquisitions (2025). **6.** KSA PDPL (SDAIA); NCA Essential Cybersecurity Controls; EU GDPR Art. 83.