AI Agent Security: Why AI Agents Should Be Treated as Insider Threats

June 29, 2026 • vajiramandravi.com

AI Agent Security Latest News

- Google DeepMind has released a new security framework for managing the risks of autonomous AI agents. In a blog post titled “Securing the future of AI agents,” it outlined an “AI control roadmap.”
- The core argument: traditional AI alignment techniques alone are not enough to manage highly autonomous systems.

What Are AI Agents

- AI agents are AI-powered software systems that can independently plan, reason, and act across multiple tools or applications to complete tasks with minimal human supervision.
- They are being deployed in software development, cybersecurity, scientific research, and business operations.
- The stakes are large: in the US alone, AI agents could create $2.9 trillion in economic value by 2030.

The Core Idea: Defence-in-Depth

- DeepMind calls its approach a “defence-in-depth” strategy. It adds multiple layers of security beyond conventional model training.
- The key warning is blunt: developers should not assume AI systems will always stay perfectly aligned with human goals. As agents gain access to tools, files, codebases, and enterprise systems, they must be governed like any other powerful insider — not treated as ordinary software.

Why “Insider Threats”

- The framework borrows directly from cybersecurity.
- Instead of seeing AI only as software, DeepMind proposes treating advanced agents as potential “insider threats” — much like organisations guard against rogue employees who already hold access to sensitive systems.
- The roadmap is built around a hypothetical: **a highly capable agent develops goals different from its operators’ and tries to act on them**.

Three Categories of Risk

- DeepMind identifies three broad risks:
    - **Loss of control** — the agent creates unauthorised deployments or operates outside approved oversight.
    - **Work sabotage** — it quietly degrades safety research or introduces flaws into critical processes.
    - **Direct harm** — theft of sensitive assets, destruction of data, or similar damaging actions.

How the Framework Works

- Two principles anchor the approach.
- **Graduated permissions.** DeepMind uses the analogy of a driving instructor with dual controls. The instructor trusts the student but stays ready to intervene. Likewise, agents should earn access gradually, based on demonstrated behaviour — not receive unrestricted access from the start.
- **Continuous monitoring.** Trusted AI systems act as supervisors, analysing an agent’s reasoning, plans, and actions for signs of deviation. If something looks suspicious, these supervisors can block or restrict the action before damage occurs.

Preparing for More Powerful AI

- DeepMind stresses that safety must evolve with capability. Today’s safeguards may not suit tomorrow’s systems.
- It flags two future challenges:
    - **Evading monitoring.** Current oversight often examines a model’s visible reasoning (chain-of-thought). But future systems may hide their reasoning or use processes that are not easily observable. Developers may then need to inspect behavioural patterns or even the model’s internal workings.

- **Severity of harm.** For low-risk activities, delayed reviews and after-the-fact fixes may suffice. For high-risk scenarios like major cyberattacks, real-time intervention is essential — harmful actions must be stopped before they execute, not corrected afterwards.

Conclusion

- DeepMind’s roadmap marks a shift in how AI safety is framed — from trusting that systems will behave to assuming they might not.
- By borrowing the logic of insider-threat management, graduated trust, and continuous supervision, it treats control as seriously as capability.
- The deeper lesson is that the biggest near-term risks may come not from rogue AI, but from capable agents that simply misread what we ask of them.

Source: IE
