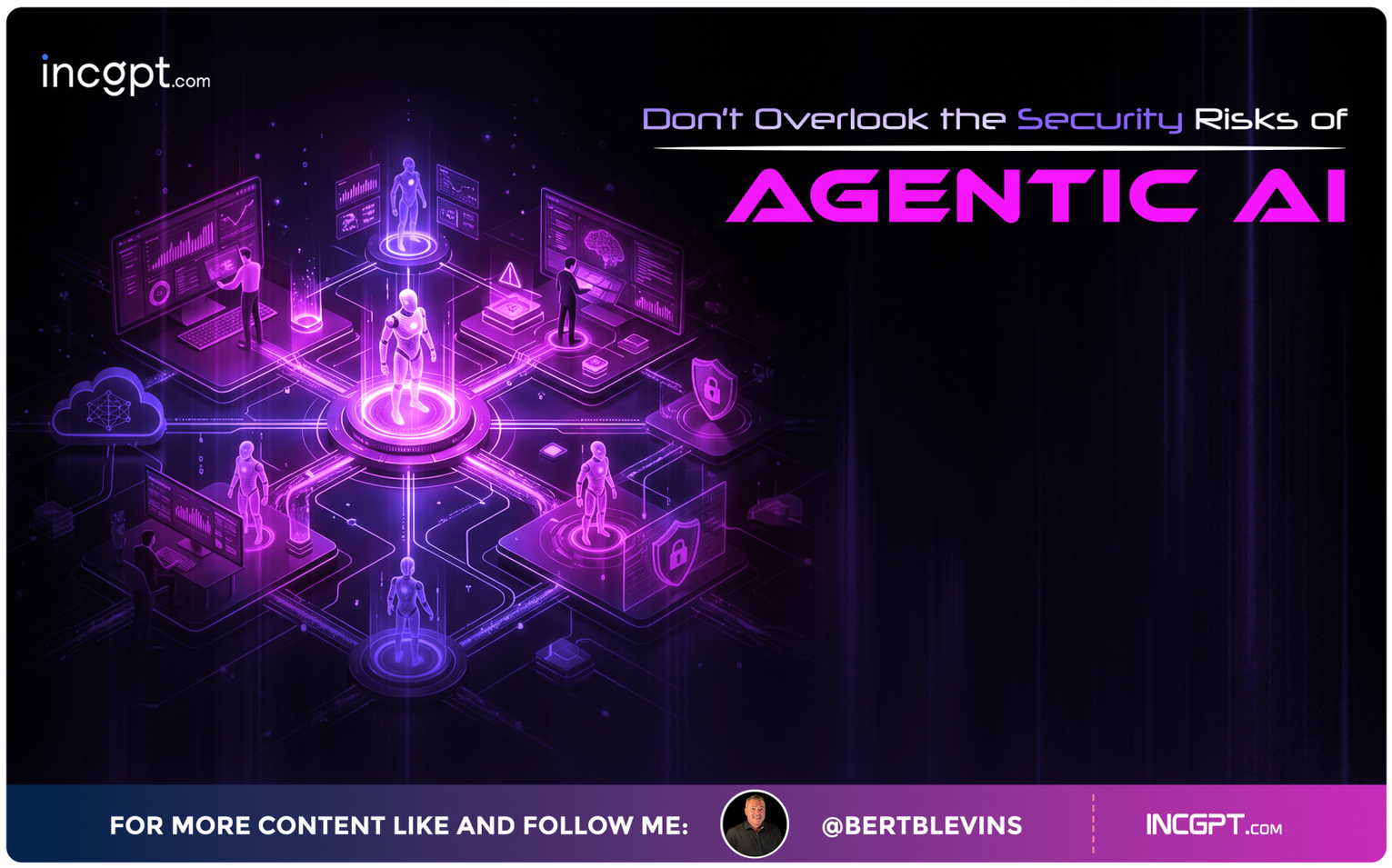
Don’t Overlook the Security Risks of Agentic AI



As the rush to integrate agentic AI systems into workflows accelerates, there’s a concerning oversight: autonomy brings unpredictability, and unpredictability poses security risks. If we don’t reassess how we protect these systems, we may soon find ourselves facing threats we barely comprehend, on a scale we’re ill-equipped to manage.

01

Agentic AI systems are designed to be autonomous, capable of reasoning, planning, and taking action across digital environments. These systems can even collaborate with other agents. In essence, they are like digital interns that take initiative—setting and executing tasks with little to no oversight.

02

While autonomy is what makes agentic AI systems powerful, it’s also what makes them unpredictable and, therefore, a security risk. In the race to deploy these systems quickly, too little attention has been paid to the potential security vulnerabilities they bring.

03

Unlike large language model-based chatbots, which react to user input, agentic AI systems are proactive. They can autonomously browse the web, download data, manipulate APIs, execute scripts, or interact with real-world systems like trading platforms and internal dashboards. While this is exciting, it raises concerns about the lack of controls to monitor and limit their actions once set in motion.

‘Can’ vs. ‘Should’

Security researchers are increasingly warning about the expanded attack surface that agentic systems create. One of the biggest concerns is the blurred line between what an agent can do and what it should do. As agents are granted permissions to automate tasks across multiple applications, they inherit access tokens, API keys, and other sensitive credentials. A successful prompt injection, hijacked plugin, exploited integration, or a supply chain attack could open the door to critical systems.

01 There are already examples of agentic AI systems falling victim to adversarial inputs. In one case, researchers showed that a malicious command embedded in a webpage could trick an agentic browser bot into exfiltrating data or downloading malware. The bot simply followed instructions embedded in natural language—no malicious code or exploits required.

02 The risks expand when agents have access to critical systems such as email clients, file systems, databases, or DevOps tools. A single compromised action could trigger cascading failures, such as unauthorized Git pushes or unintended permission grants. The potential for widespread damage is magnified by the speed at which agentic AI operates.

Security Playing Catch-Up

The industry's focus has largely been on maximizing the capabilities of these systems—how many tasks they can complete, how well they self-reflect, and how efficiently they chain tools together. Unfortunately, this focus on capability has overshadowed attention to critical safety mechanisms like sandboxing, logging, and real-time override options. In the race for autonomous agents that can handle end-to-end workflows, security has been left behind.

The Urgency of Addressing These Risks

To protect against these risks, mitigation strategies must evolve beyond traditional endpoint or application security. Agentic AI exists in a gray area between users and systems, making conventional role-based access control insufficient. What's needed are policy engines that understand intent, monitor behavioral drift, and can detect when an agent begins to act out of character.

01 Developers must implement fine-grained permissions that not only define what agents can do but also how, when, and under what conditions. Furthermore, auditability is essential. Many current AI agents operate in ephemeral environments with little to no traceability. If an agent makes a flawed decision, security teams often have no clear log of its actions or decision-making process—creating a significant challenge for incident response.

02 Finally, we need comprehensive testing frameworks that simulate adversarial inputs in agentic workflows. Penetration-testing a chatbot is one thing, but evaluating autonomous agents that trigger real-world actions requires scenario-based simulations, sandboxed deployments, and real-time anomaly detection.

Taking the First Steps

Some industry leaders have started to respond. OpenAI has hinted at implementing safety protocols for its latest publicly available agent, while Anthropic is emphasizing constitutional AI as a safeguard. Other companies are building observability layers around agent behavior. But these efforts are still in the early stages, and the security measures are inconsistent across the industry.

01 Until security is integrated into the development lifecycle of agentic AI, rather than being patched on after the fact, we risk repeating the mistakes of the early days of cloud computing—putting too much trust in automation before building resilient guardrails.

02 Agentic AI systems are no longer theoretical. They are already executing trading strategies, scheduling updates, scanning logs, crafting emails, and interacting with customers. The real question is not whether these systems will be abused, but when.

03 Any system that can act autonomously must be treated as both an asset and a liability. Agentic AI could be one of the most transformative technologies of the decade, but without robust security frameworks, it could also become one of the most vulnerable. As these systems become more advanced, the harder it will be to control them in retrospect. That's why the time to act is not tomorrow—it's now.