Is AI Agent the new security 'nightmare'?

AI agents offer many benefits but also pose new security risks. Learn about four major threats businesses need to be aware of.

2026 is considered the year of the explosion of automated AI agent systems. We are witnessing a strong shift from purely responsive chatbots to proactive AI agents capable of planning and reasoning, often integrated with large language models (LLMs) or RAG systems.

This shift is changing the entire cybersecurity landscape. The reason is simple: AI agents not only answer questions but can also take action. They can send mass emails, manipulate databases, and interact with internal systems or external applications without direct human intervention. This significantly increases the complexity of security.

So, will AI agents really become the next security nightmare? Here are four core issues that are causing concern among experts.

1. Shadow AI and the excessive freedom of AI agents.

Shadow AI refers to the deployment of AI tools without formal oversight or control. This is one of the biggest risks today.

A prime example is OpenClaw—an open-source AI agent tool that can control personal or work accounts with very few restrictions. According to reports from early 2026, many OpenClaw systems were exposed on the internet without security authentication, allowing hackers to gain complete control of the servers.

This raises a crucial question for businesses: should employees be allowed to freely deploy AI agents without IT supervision?

2. Supply chain risks

AI agents often rely on plugins, extensions, or third-party skills to connect to external systems via APIs. This creates a new software supply chain but also carries significant risks.

Malicious plugins can be disguised as productivity tools. Once installed, they can access sensitive data, execute code remotely, or install malware without the user's knowledge.

3. New forms of attack

The OWASP security report indicates that 2026 will see the emergence of many new risks related to AI agents. One of these is "Agent Goal Hijack," which involves hackers changing the AI's target through hidden instructions on the web.

Additionally, AI agents typically store both short-term and long-term data between sessions. This helps the AI ?? operate more efficiently but also makes them vulnerable to manipulation if the data is contaminated.

Other risks include over-access and supply chain vulnerabilities, which have already been mentioned above.

4. Lack of a 'circuit breaker' mechanism when AI malfunctions.

Traditional security systems are no longer effective enough in environments with multiple AI agents. Because these agents operate at very high speeds, a small vulnerability can spread throughout the entire system in just milliseconds.

Many businesses currently lack a 'circuit breaker' mechanism to detect and stop AI when it behaves abnormally. This significantly increases the risk.

How can we reduce security risks?

Security experts argue that the most important principle is: you can't protect what you can't see. Therefore, businesses need to change their security strategy when deploying AI agents. Some solutions include:

1. Enhance runtime monitoring capabilities.
2. Apply the principle of least access.
3. View the AI ??agent as a unique identity within the system.
4. Establish a confidence score for each agent.

If managed properly, AI agents don't necessarily become a security nightmare. On the contrary, they can become powerful tools to help businesses increase productivity while maintaining security.

AI agents are opening up many opportunities but also presenting new security challenges. Organizations need to build appropriate governance frameworks to leverage the benefits of AI without compromising security.

If implemented correctly, AI agents are not a threat, but a crucial foundation for the future of businesses.