

# Microsoft says enterprise AI agents need Guardian Agents to watch every action

Source copy kept by wheresthe.ai for the story "Microsoft says enterprise AI agents need Guardian Agents to watch every action" (<https://wheresthe.ai/d/intl-microsoft-describes-guardian-agents-to-monitor-enterpri-2026-10/>).

Source: <https://blogs.microsoft.com/blog/2026/10/19/securing-the-agentic-insider>

Text saved 19 Oct 2026 (the page as published).

SHA-256 of the saved text: 260482f7a1fc4d737b4f7e7c73991a4d8380ddf36e1693637f2b5049f4f5dd5e

---

In the last few months, concerns have arisen about AI safety as a growing number of incidents were publicly disclosed. These incidents exposed how powerful models have become and how they represent a significant threat if not handled well. And with the global race on innovation, the challenge will compound as models become even more powerful.

While most have been worried about threat actors using powerful models to attack, the focus has been less on agents within our own walls. Since we were the ones who deployed agents internally for our companies, we implicitly trusted our agents.

Unfortunately, frontier models, regardless of their origin, cannot be trusted. This is not because models are inherently malicious. Models have been trained to attain results, which often leads to reward hacking, and models are not optimized around organizational norms or judgment. Agents are unpredictable. They will adapt as the context changes, may not behave as intended and will collaborate and collude with other agents to reach their objectives, without needing to be compromised.

Agents have been unleashed in the enterprise, at scale, and their numbers are growing rapidly. These agents are frequently overprivileged, have access to data, tools and systems and can take actions and make decisions in critical processes. They have become active participants in the enterprise.

They are agentic insiders.

Monitoring and securing those agents must become continuous. We need to understand in real time what the agents are doing and what risks those actions create, and we need to adjust their ability to act accordingly.

To manage and control these insiders, we will need Guardian Agents — an established industry term — that watch and moderate every action of every agent, before it happens, 24/7.

An agent brings together a model, a harness that orchestrates its work, an action space of tools and resources and a runtime in which it executes. Securing agents requires protection across multiple layers: identity, permissions, data, endpoints, network, cloud and runtime.

In July, we announced Project Perception, an agentic security system that uses our new cyber stack and teams of red, blue and green agents to help defenders act with speed, while defending against AI driven attacks from outside of the enterprise.

The same cyber stack is also directly applicable to managing and governing the agentic insiders that run in the enterprise.

Project Perception can be leveraged to contain agentic insiders, implementing deterministic controls on non-deterministic intelligence:

**Sensors/Signals/Actuators:** bring together signals across the various sensors (identity, data, cloud, endpoint, network, runtime) to discover agents in the enterprise and bring them under management by assigning an identity, container and set of policy and permissions.

**Context:** use those signals to compute, dynamically, the risk associated with what the agents are attempting to do.

**Harness/agents/models:** based on dynamic agent risk, a team of Guardian Agents, with green and blue agents, actively and continually supervise every action of every agent and surgically respond as the risk profile increases, in real time.

This is where agentic security changes the equation for security practitioners.

Project Perception can supervise in real time while providing traceability for practitioners to examine what happened, understand why the risk emerged and apply the judgment needed to update their security system accordingly. Agentic security provides speed . Practitioners provide judgment . That is how defenders begin operating at the same speed as the systems they are protecting against.

Today, security systems can reliably observe the actions of the agents. Those signals are essential and tell us what we need to contain. Our security controls should increasingly be able to ask not only “What is happening?” but also “Why is it happening?” Is this consistent with the intent of the task?” This will require greater transparency into the model itself, not only the input and output tokens, but the thinking and chain of thought tokens. Only then can we truly understand whether the model’s intent is aligned with the purpose of the agent. Model transparency will not replace containment. It will make containment smarter and stronger . And it will allow us to better inform accountability and traceability. This needs to be the next step in securing agentic insiders.

With Project Perception, you can get started today on the journey to secure your agentic insiders . There should not be a choice between innovation and safety.

Hayete Gallot leads Microsoft’s work to help organizations operate securely in an AI-driven world. Her scope includes identity, threat protection, compliance and data security at global scale.

Tags: Agentic AI , AI , Project Perception , Security