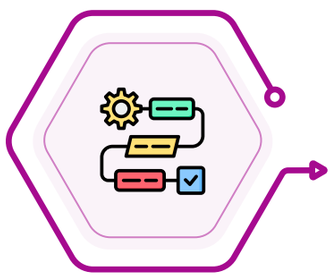
Microsoft Builds a Security Cage for AI Agents — and OpenAI and Nvidia Are Already Inside

Buried inside a crowded slate of developer announcements at Build this year, Microsoft introduced what may be its most far-reaching platform decision of the event: Microsoft Execution Containers, or MXC. It is a policy-driven execution layer baked directly into Windows, and it sets out to answer the question every security team has been quietly dreading — how do you let an autonomous AI agent run loose on a corporate machine without handing it the keys to everything?

The premise is simple and, for anyone who manages identity and access, overdue. Instead of trusting an agent to behave, MXC lets developers and IT administrators declare exactly what an agent is permitted to touch, and then enforces those boundaries at runtime through the operating system kernel itself. The rules do not live in the application. They live in the OS, beneath the agent, where the agent cannot reach them.

Not a product you buy

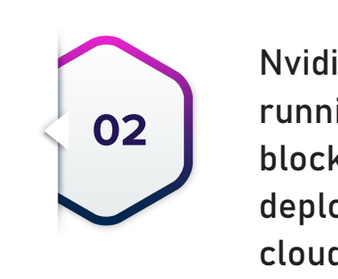
MXC is not something an enterprise purchases off a price sheet. It ships as an SDK and a policy model — a foundational primitive embedded in both Windows and the Windows Subsystem for Linux. Microsoft describes what it offers as a "composable sandbox spectrum," and that phrase is worth unpacking, because it is the heart of the design.



At the lightest end of the spectrum sits simple process isolation, the kind already in use by the GitHub Copilot command-line interface. From there, organizations can dial the containment up as the risk profile demands: micro virtual machines, Linux containers, and ultimately full cloud instances. The point is that a single policy framework governs all of it. A security team does not have to bolt together separate tools for a lightweight coding helper and an always-on autonomous worker. The same controls scale across the whole range.

Why identity teams should care

For anyone working in privileged access, the most consequential part of MXC has nothing to do with sandboxing performance. It is attribution. Every action an agent takes can be tied to a specific identity, and the same policy infrastructure that already governs hundreds of millions of Windows devices extends to govern the agent.



That matters enormously for regulated sectors. In financial services, healthcare, and government, the ability to produce an audit trail that cleanly separates human actions from agent actions on the same machine is shifting from a convenience to a compliance expectation. When an agent and its operator share a workstation, “the system did it” is not an acceptable answer to an auditor. MXC is designed to make every agent action accountable to an identity and containable through policy — which is precisely the architecture that could finally push AI agents out of pilot purgatory and into production.

The partners are the real signal

Platform announcements at developer conferences tend toward the aspirational. What separates MXC from vaporware is the specificity of the partners already building on it. Microsoft named five: OpenAI, Nvidia, Manus, Nous Research (the team behind the Hermes agent), and the open-source OpenClaw project.

01

OpenAI’s involvement is the headline. A member of its technical staff framed the collaboration as a way to explore safer patterns for agents to generate and execute code, pairing the capabilities of Codex — OpenAI’s code-generation agent — with the controlled execution environment MXC provides. Read between the lines and the implication is striking: MXC could become the default place where one of the industry’s most anticipated agent products actually runs.

02

Nvidia, for its part, is bringing its OpenShell runtime to Windows on top of MXC. OpenShell is Nvidia’s open-source framework for running autonomous agents in sandboxed environments with kernel-level isolation, governed by declarative YAML policies that block unauthorized file access, credential theft, and data exfiltration. Layered onto MXC, it gives developers a packaged way to deploy always-on agents while handing IT consistent visibility and policy enforcement across devices, virtual machines, and the cloud. Manus, the Chinese-born agent startup that went viral earlier this year, is integrating as well.

The bottom line

The agent era has a trust problem. Autonomous software that reads files, installs packages, calls APIs, writes its own code, and runs for hours without supervision is enormously useful and equally dangerous — and most enterprises have been stuck running it in cautious pilots precisely because the containment story was missing. MXC is Microsoft’s attempt to supply that missing layer at the level where it is hardest to bypass: the operating system.

For identity and security practitioners, the takeaway is clear. The conversation is moving from whether agents can be deployed to how their access is scoped, attributed, and audited. MXC plants that conversation firmly in the same policy plane that already governs the enterprise endpoint — and with OpenAI and Nvidia already on board, it is unlikely to stay a Build footnote for long.