PANOP

PANOP RESEARCH · AWARENESS BRIEFING

Agents, MCP and the New Perimeter

An awareness briefing on autonomous AI agents, the Model Context Protocol and the tool-calling estate now spreading through enterprise environments: what these systems are, the risks they introduce, and the governance, controls, frameworks and certifications needed to hold them to account.

EDITION AUGUST 2026

PANOP RESEARCH · PANOP.IO

PANOP

01 · EXECUTIVE SUMMARY

The model stopped answering and started acting

An AI agent is a non-deterministic actor: it decomposes a goal, selects its own tool calls and executes them against production systems under delegated credentials. The Model Context Protocol standardises that access — and with it the blast radius. The result is a privileged identity class no IAM, change-control or DLP boundary currently sees.

THE SHIFT IN ONE PARAGRAPH

Until recently an enterprise language model was a text interface: it produced output and a human decided what to do with it. Agentic systems remove that step. The model selects a tool, invokes it, reads the result and chooses the next action, in a loop, at machine speed. MCP standardised the model-to-tool connection, which made the pattern deployable at scale and turned thousands of independently written tool servers into one shared attack surface.

5,000+

Servers in the official MCP registry by mid-2026

Registry, 2026

40+

CVEs against MCP SDKs disclosed Jan–Apr 2026

Multiple advisories

200,000

MCP instances assessed vulnerable, April 2026

OX Security, 2026

150M+

Package downloads in the affected supply chain

CSA, 2026

WHAT LEADERSHIP HAS TO DECIDE

01 Do you know what agents you run?

Most organisations cannot produce a list of deployed agents, the tools each can call, or the credentials those tools hold. Inventory is the precondition for every other control.

02 Who is accountable when an agent acts?

An agent is a non-human identity making privileged decisions. Ownership, approval thresholds and a named accountable human must be assigned per agent, not per platform.

03 Is a third-party MCP server a supplier?

It executes code inside your trust boundary with your credentials. Treat server adoption as vendor onboarding, with review, pinning and provenance, not as installing a plug-in.

04 Where does a human stay in the loop?

Autonomy needs tiers. Read-only retrieval, reversible writes, and irreversible or financial actions warrant different levels of confirmation and out-of-band verification.

05 Can you prove any of it after the fact?

Prompts, tool calls, parameters and outputs are the audit trail of an agentic system. If they are not logged and retained, incidents cannot be reconstructed or explained to a regulator.

> **The governance gap.** Existing controls assume a human actor and a deterministic system. Agentic AI breaks both assumptions at once.

Sources Anthropic, Introducing the Model Context Protocol, November 2024. NSA / CISA, Cybersecurity Information Sheet: Model Context Protocol (MCP) Security, June 2026. Cloud Security Alliance, MCP Security Crisis: Systemic Design Flaws in AI Agent Infrastructure, May 2026. OX Security advisory, April 2026. *Full citations on page 9.*

PANOP

02 · WHAT WE ARE ACTUALLY TALKING ABOUT

Five layers, and the trust boundary is in the wrong place

Agentic systems are not a single product. They are a chain of five components, each with a different owner, and the security properties of the whole chain are set by its weakest link.

REFERENCE ARCHITECTURE OF AN AGENTIC SYSTEM

01

Trigger

Human prompt, schedule, webhook or another agent. Increasingly not a person at all.

02

Model and planner

The LLM decomposes the goal, selects tools and interprets results. Non-deterministic by design.

03

Agent runtime

Memory, state, loop control and the credential store. Where autonomy is bounded, or is not.

04

MCP client and servers

One standard interface to tools. Server code, schemas and descriptions enter model context.

05

Tools and data

Repositories, mailboxes, ticketing, cloud APIs, payment rails. Real and often irreversible effect.

Trust boundary. Tool descriptions, schemas, outputs and credentials all sit inside it. Anything the model reads can influence what it does next.

PLAIN-LANGUAGE GLOSSARY

AI agent

A language model given tools, memory and a control loop, permitted to take actions toward a goal with limited or no human confirmation at each step.

Model Context Protocol (MCP)

An open standard, released by Anthropic in November 2024, for exposing tools, resources and prompts to a model through one uniform interface. By mid-2026 it is supported across the major assistants and development environments.

MCP server

A process that advertises tools to an agent and executes them. It may be first-party, open-source or vendor-hosted, and it runs with real credentials against real systems.

Tool poisoning

Malicious instructions hidden in a tool's description, schema or output, which the model reads as trusted context and then acts upon.

Non-human identity

The credential an agent authenticates with. It has no line manager, no joiner-mover-leaver process, and frequently no expiry.

Why it matters. Standardisation means one exploit now applies across thousands of deployed servers, not one bespoke integration.

Sources *Anthropic, Model Context Protocol specification and registry, 2024–2026. Microsoft Security, The state of MCP security in 2026, June 2026. NSA / CISA, Cybersecurity Information Sheet: MCP Security Design, June 2026. Li and Gao, A First Look at the Security Issues in the MCP Ecosystem, arXiv 2510.16558, rev. April 2026.*

PANOP

03 · THE RISK SURFACE

Seven attack classes that did not exist three years ago

These are not theoretical. Each class below has been demonstrated against production systems, and most were disclosed with a CVE or a named research write-up between mid-2025 and mid-2026.

SEVEN CLASSES, AND ONE COMBINATION THAT MATTERS MOST

Prompt injection, direct and indirect

Untrusted content that the model reads becomes instruction. It is the common thread running through nearly every agentic attack observed to date.

Ranked LLM01, OWASP Top 10 for LLM Applications

Tool poisoning and tool shadowing

Hostile instructions planted in a tool description or schema, or a rogue server silently overriding a legitimate tool's behaviour.

Invariant Labs, GitHub MCP server, May 2025

Confused deputy and token passthrough

A server acts with its own broad privileges on behalf of a user who does not hold them, or a proxy is tricked into issuing a valid authorisation code.

CVE-2026-13341, Kong Konnect MCP server

Rug pulls and supply-chain compromise

A vetted server changes behaviour in a later version. Install-time review does not catch what ships in the next release.

Nx npm compromise: circa 2,180 tokens, 20,000 files

Rogue server registration

Weak registry vetting and ownership checks allow adversarial or hijacked servers to be adopted by hosts as though they were legitimate.

Li and Gao, arXiv 2510.16558, April 2026

Remote code execution in the plumbing

Unauthenticated command injection, DNS rebinding and localhost-binding flaws in clients, SDKs and inspectors rather than in the model itself.

CVE-2025-6514, CVSS 9.6, 437,000 environments

Excessive agency

The agent is able to do far more than the task requires because scopes are ambient and session-wide rather than scoped per call.

Tracked as LLM06, OWASP

The lethal trifecta

External content, access to private data, and outbound network reach. An agent holding all three can be induced to exfiltrate, whatever the prompt said.

Simon Willison, 2025; widely adopted framing

The pattern. Almost none of these are model failures. They are identity, supply-chain and authorisation failures in the layer around the model.

Sources *OWASP Top 10 for LLM Applications and the OWASP MCP Top 10, 2025. OWASP Agentic AI Top 10 (emerging). Wiz, Understanding Model Context Protocol Security, 2026. Invariant Labs, GitHub MCP server exfiltration, May 2025. Backslash Security, NeighborJack, April 2025. NVD: CVE-2025-6514, CVE-2025-49596, CVE-2026-13341.*

PANOP

04 · WHY EXISTING GOVERNANCE BREAKS

Every control you own assumes a human clicked the button

The controls are not wrong. They are scoped to an actor that is provisioned, trained, supervised and accountable. An agent is none of those things, so each control needs an explicit extension.

CONTROL DOMAIN	THE CLASSIC ASSUMPTION	WHAT AN AGENT DOES INSTEAD
Identity and access	Users are provisioned, reviewed and deprovisioned through a joiner-mover-leaver process.	Agents authenticate with long-lived tokens and service accounts that no HR process ever touches.
Authorisation	Permissions are granted per role and checked once at the session boundary.	Ambient, session-wide scopes let a single poisoned input reach every tool the session can call.
Change management	Code changes pass peer review and testing before they reach production.	A tool description or a server update alters system behaviour with no code change on your side.
Third-party risk	Suppliers are onboarded, assessed, contracted and periodically reassessed.	MCP servers are added by developers in minutes and then execute inside the trust boundary.
Data classification	Data flows are mapped between a known set of systems and interfaces.	The agent assembles context across systems at runtime, creating flows no diagram anticipated.
Logging and audit	Logs record who did what, to which record, and when.	Without prompt, tool-call and parameter logging there is no record of why an action was taken.
Awareness and training	Staff are trained to recognise a suspicious request and not to click the link.	The agent has no scepticism and no training. It reads the instruction and complies with it.

> **The governance move.** Do not build a parallel AI programme. Extend the seven domains you already run and give each an agentic acceptance criterion.

Sources *Panop analysis mapped against NIST AI RMF 1.0 (NIST AI 100-1), ISO/IEC 27001:2022 Annex A, ISO/IEC 42001:2023 and the CSA AI Controls Matrix. Microsoft Security, The state of MCP security in 2026. NSA / CISA, MCP Security Design, June 2026.*

PANOP

05 · PRACTICES, FRAMEWORKS AND CONTROLS

Four functions, one control set, no second programme

There is no shortage of frameworks. The task is not to adopt all of them but to pick one spine, map the rest onto it, and turn the result into control statements an auditor can test.

THE SPINE: NIST AI RMF 1.0 (AI 100-1)

GOVERN

Accountability, policy and culture. Name an owner for every agent and set autonomy tiers.

MAP

Context and inventory. Enumerate agents, tools, servers, data reached and credentials held.

MEASURE

Test and monitor. Red-team for injection, track tool-call anomalies and blast radius.

MANAGE

Respond and improve. Kill switches, revocation paths and post-incident reconstruction.

THE FRAMEWORK MAP

FRAMEWORK	WHAT IT GIVES YOU	HOW TO USE IT
NIST AI RMF 1.0	Voluntary, sector-agnostic risk framework built on Govern, Map, Measure, Manage.	Adopt as the operating spine and the language for board-level reporting.
ISO/IEC 42001:2023	The first certifiable AI management system standard, structured like ISO/IEC 27001.	Certify the organisation. Gives procurement a recognisable assurance artefact.
OWASP LLM, Agentic and MCP Top 10	Practitioner threat taxonomies covering injection, excessive agency and MCP-specific risk.	Drive threat modelling and write security test cases from it.
MITRE ATLAS	Adversary tactics and techniques observed against AI-enabled systems.	Structure red-teaming, purple-team exercises and detection engineering.
CSA AI Controls Matrix	A technical control overlay mapped across cloud and AI risk domains.	Derive auditable control statements and evidence requirements.
NSA / CISA MCP guidance	Government hardening guidance specific to MCP design and deployment.	Use as the baseline configuration checklist for every MCP server.

EIGHT CONTROLS THAT ACTUALLY REDUCE RISK

- Per-tool, least-privilege scopes, deny by default
- Pinned server versions with provenance and signature checks
- Human confirmation gates on irreversible and financial actions
- Egress control and network isolation for agent runtimes
- Short-lived tokens; never pass user credentials through
- Allow-listed servers from an internal registry, not ad-hoc install
- Full logging of prompts, tool calls, parameters and outputs
- Continuous red-teaming for injection and tool poisoning

Sources *NIST, Artificial Intelligence Risk Management Framework AI RMF 1.0, NIST AI 100-1, January 2023. ISO/IEC 42001:2023. OWASP Top 10 for LLM Applications, OWASP MCP Top 10 and Agentic AI Top 10, 2025-2026. MITRE ATLAS. Cloud Security Alliance, AI Controls Matrix. NSA / CISA, MCP Security Design, June 2026.*

PANOP

06 · REGULATION AND CERTIFICATION

A certificate proves the house is managed, not that the product is safe

Regulation and certification answer different questions. The EU AI Act regulates systems placed on the market; ISO/IEC 42001 certifies the organisation managing them. Buyers increasingly ask for both.

EU AI ACT: THE APPLICABLE TIMELINE

WE ARE HERE

1 Aug 2024	2 Feb 2025	2 Aug 2025	2 Aug 2026	2 Dec 2027	2 Aug 2028
Regulation (EU) 2024/1689 enters into force.	Prohibited practices and the Article 4 AI literacy duty apply.	Obligations for general-purpose AI model providers apply.	Article 50 transparency duties apply; the Commission gains GPAI enforcement powers.	Annex III standalone high-risk duties, moved back by the Digital Omnibus.	High-risk AI embedded in regulated products such as lifts and toys.

THE CERTIFICATION LANDSCAPE

ISO/IEC 42001:2023

Certifies the organisation's AI management system. Published December 2023, it has become the default procurement signal. It does not certify any individual AI system.

Maps to Articles 9-14 and 17

prEN 18286

The first quality management standard written for per-system conformity under Article 17. Still at CEN Enquiry stage, so begin the gap analysis rather than waiting for publication.

Pre-validated against Article 17

ISO/IEC 27001 and SOC 2

Still the base layer. Extend existing scope statements to cover agent runtimes, MCP servers and non-human identities instead of certifying an AI silo separately.

Extend scope, do not duplicate

WHAT A CERTIFICATE DOES NOT DO

- **No presumption of conformity.** An ISO/IEC 42001 certificate is not a legal defence under the AI Act.
- **Organisation, not product.** The Act regulates systems placed on the market; the standard certifies the management system around them.
- **Auditor interpretation varies.** Assessed against ISO 42001 alone, coverage of the thirteen Article 17 elements is a matter of judgement.
- **Scope statements beat logos.** Ask which systems, which sites and which dates a certificate actually covers.

Penalties. Up to €35 million or 7% of worldwide annual turnover for prohibited practices, €15 million or 3% for other high-risk breaches, and €7.5 million or 1.5% for supplying incorrect information to authorities.

Sources *Regulation (EU) 2024/1689. European Commission implementation timeline and the Digital Omnibus political agreement of 7 May 2026. ISO/IEC 42001:2023. prEN 18286 (CEN Enquiry stage). Cloud Security Alliance, EU AI Act, prEN 18286 and ISO 42001, April 2026. Note: high-risk dates have moved once and may move again.*

PANOP

07 · WHAT TO DO NEXT

Ninety days to provable control

This is achievable in a quarter — but only if agents, their credentials and every tool they can reach land in one inventory, prioritised by what is actually exploitable and revalidated as it changes. That is the capability to insist on, and the evidence a third party would accept.

A 90-DAY SEQUENCE

DAYS 0-30 · SEE

- Publish an interim rule: no new MCP server without review.
- Inventory every agent, tool and server in use, developer machines included.
- List the credentials each agent holds and the systems it can reach.
- Name an accountable owner for each agent.
- Switch on prompt and tool-call logging wherever the platform allows.
- Flag any agent holding all three legs of the lethal trifecta.

DAYS 31-60 · CONTAIN

- Replace long-lived tokens with short-lived, per-tool scoped credentials.
- Stand up an internal registry of approved servers with version pinning.
- Add human confirmation gates to irreversible and financial actions.
- Isolate agent runtimes and apply egress control.
- Fold agents into joiner-mover-leaver and access review cycles.
- Extend incident response with an agent kill switch and revocation path.

DAYS 61-90 · PROVE

- Map your controls to NIST AI RMF functions and ISO/IEC 42001 clauses.
- Red-team the three highest-privilege agents for injection and poisoning.
- Add agentic criteria to procurement and third-party questionnaires.
- Brief the board on residual risk and agreed autonomy tiers.
- Decide the certification path and draft the scope statement.
- Set a quarterly reassessment; the estate changes every month.

FIVE QUESTIONS FOR THE NEXT BOARD MEETING

- Where can an agent commit this company — spend, contracts, customer data — without a human signing off?
- Who is accountable when an agent causes loss, and is that named in our risk register?
- If an agent were manipulated tomorrow, could we prove to a regulator or insurer what it did?
- What is our exposure through AI tools that teams adopted without procurement or security review?
- Are we on record with a position on the EU AI Act duties that apply from 2 August 2026?

> **The test.** If you cannot answer all five in one meeting, the inventory work in the first thirty days is where to start.

Sources *Panop advisory practice, synthesised against NIST AI RMF 1.0, ISO/IEC 42001:2023, the NSA / CISA MCP Cybersecurity Information Sheet of June 2026, Microsoft Security guidance on MCP hardening, and the OWASP MCP and Agentic AI Top 10 control recommendations.*

PANOP

08 · REFERENCES AND METHOD

Every claim traces back to a public source

Use this page to check anything in the briefing, and to hand colleagues the primary material rather than the summary.

STANDARDS AND FRAMEWORKS

NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, January 2023.

ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system, December 2023.

ISO/IEC 27001:2022, Information security management systems — Requirements.

prEN 18286, Quality management system for AI systems, CEN, Enquiry stage, 2026.

OWASP, Top 10 for Large Language Model Applications, 2025; MCP Top 10, 2025; Agentic AI Top 10 (emerging).

MITRE ATLAS, Adversarial Threat Landscape for Artificial-Intelligence Systems.

Cloud Security Alliance, AI Controls Matrix (AICM).

GOVERNMENT AND REGULATORY

Regulation (EU) 2024/1689, Artificial Intelligence Act, in force 1 August 2024.

European Commission, AI Act implementation timeline; Digital Omnibus political agreement, 7 May 2026.

NSA and CISA, Cybersecurity Information Sheet: Model Context Protocol (MCP) Security Design, June 2026.

European Commission, General-Purpose AI Code of Practice, 2025.

USING THIS BRIEFING

Circulate sections 01 to 04 for general awareness; sections 05 to 07 are written for security, risk and legal owners.

The 90-day sequence in section 07 is designed to be lifted directly into a workplan.

Re-verify every figure and date before any external use. This material dates quickly.

THREAT RESEARCH AND INCIDENTS

Anthropic, Introducing the Model Context Protocol, November 2024; MCP registry, 2026.

Microsoft Security, The state of MCP security in 2026, June 2026.

Wiz, Understanding Model Context Protocol Security, 2026.

Cloud Security Alliance, MCP Security Crisis: Systemic Design Flaws in AI Agent Infrastructure, May 2026.

OX Security, MCP architectural flaw advisory, April 2026.

X. Li and X. Gao, A First Look at the Security Issues in the Model Context Protocol Ecosystem, arXiv:2510.16558, revised April 2026.

Invariant Labs, GitHub MCP server data exfiltration, May 2025.

Backslash Security, NeighborJack: publicly exposed MCP servers, April 2025.

NVD: CVE-2025-6514 (mcp-remote), CVE-2025-49596 (MCP Inspector), CVE-2026-13341 (Kong Konnect), CVE-2026-33032 (nginx-ui).

METHOD AND LIMITATIONS

This is an awareness briefing, not an audit, a penetration test or a legal opinion.

Figures are reproduced as published by the cited source and were current at the edition date.

The MCP ecosystem moves quickly. Vulnerability counts, registry sizes and regulatory dates should be re-verified before any external use of this document.

EU AI Act high-risk application dates have already been amended once and may change again.

Prepared by *Panop Research, edition August 2026. This document is intended for internal awareness and briefing use. It is not legal advice and does not constitute an assurance opinion.*