AI Trust & Governance Incident Report

June 2026

AI adoption is accelerating, but so are governance, security, and operational risks. This report highlights five recent incidents that show why organizations need stronger oversight and trust-assessment frameworks. None was caused by malicious AI.

5	4	3	3	3
Incidents reviewed	Security	Governance	MCP-related	AI agent

Five incidents at a glance

01 AgentJacking targets MCP-connected AI agents

02 Frontier-model safety draws government attention

03 An AI coding agent reportedly deletes a production database

04 Large numbers of publicly exposed MCP servers found

05 Multiple MCP vulnerabilities highlight ecosystem risks

Incidents & Governance Lessons

#01 AgentJacking targets MCP-connected AI agents

Malicious instructions hidden in data an AI coding agent reads through an MCP server lead it to execute attacker-influenced actions with the user's own privileges. Because every step is authorized, the chain bypasses firewalls, EDR, and IAM.
Takeaway: Trust boundaries and runtime monitoring matter.

#02 Frontier-model safety draws government attention

A leading lab called for mandatory third-party testing of advanced models while authorities issued export-control directions restricting certain frontier models on national-security grounds.
Takeaway: AI capabilities often evolve faster than governance frameworks.

#03 An AI coding agent reportedly deletes a production database

An agent removed a production database and its backups in seconds using valid credentials and approved APIs after misreading its environment. The failure was access control, not a malicious model.
Takeaway: Human approval workflows and permission controls remain critical.

#04 Large numbers of publicly exposed MCP servers found

Internet-wide scans catalogued tens of thousands of reachable MCP servers, with a large share running with no authentication and many relying on static, long-lived API keys.
Takeaway: Organizations are adopting MCP faster than they are governing it.

#05 Multiple MCP vulnerabilities highlight ecosystem risks

Dozens of disclosed vulnerabilities spanned authentication, authorization, remote code execution, and information disclosure. Root causes were mostly fundamentals: missing input validation, absent authentication, and blind trust in tool descriptions.
Takeaway: MCP ecosystems require ongoing security review.

Common Themes Across All Incidents

- ! Excessive Permissions
- ! Lack of Visibility
- ! Weak Authentication
- ! Poor Governance
- ! Missing Audit Controls
- ! Limited Human Oversight

Root cause: insufficient trust & governance — not malicious AI

What Organizations Should Do Now

- ☐ Inventory AI agents
- ☐ Review MCP integrations
- ☐ Review permissions
- ☐ Enable audit logging
- ☐ Establish governance controls
- ☐ Perform trust assessments
- ☐ Conduct risk reviews

Metinc plans to publish a monthly AI Trust & Governance Incident Report. Subscribe at metinc.com to receive future editions.