

Zero Trust for MCP: Governing What AI Agents Do, Not Just What They Say

An Igris Security Whitepaper



IGRIS SECURITY
WHITEPAPER

Zero Trust for MCP

Governing What AI Agents Do

Zero trust for AI.

Every tool call and every token: governed, audited, controlled.



Your AI agent has keys to your database, your CRM, and your cloud. Who approved its last 10,000 actions?

This whitepaper is for CISOs, CTOs, and the security and platform leaders who work with them in fintech, healthtech, legal tech, and enterprise SaaS. It explains how MCP changes AI risk, sets out eight controls for governing agent actions, and maps each one to the frameworks your auditors and customers already use.

1. Executive summary

AI agents no longer just answer questions. Through the Model Context Protocol (MCP), they read files, query databases, update CRMs, and run commands, using credentials your company issued. Most security teams cannot see those actions, let alone approve them.

Traditional access control authorizes an agent once, at login. MCP needs a decision on every tool call, because the actor making that call is a language model that can be steered by the very text it reads.

This paper sets out eight controls that put one policy checkpoint in front of every MCP server. It maps them to NIST AI RMF, ISO/IEC 42001, the EU AI Act, GDPR, India's DPDP Act, and more, and gives you a quick way to score where you stand. Igris appears only in the final chapter.

2. From chat to action

MCP turned AI from something that talks into something that acts, and the security model has not caught up.

In plain terms, MCP is a standard plug. An AI agent (the client) connects to MCP servers, and each server exposes tools: query this database, create that ticket, read this folder. The server holds credentials to the real system behind it. When the model decides to call a tool, the action happens with your company's authority.

Three things make this different from any API you have secured before. The decision-maker is a language model, and it can be steered by text it reads in an email, a document, or a ticket. Servers are often granted

far broader permissions than the task needs. And new servers are easy to install, so they spread through engineering teams faster than security can review them.

FIGURE 01

Agents act with your authority, and trust breaks at four points

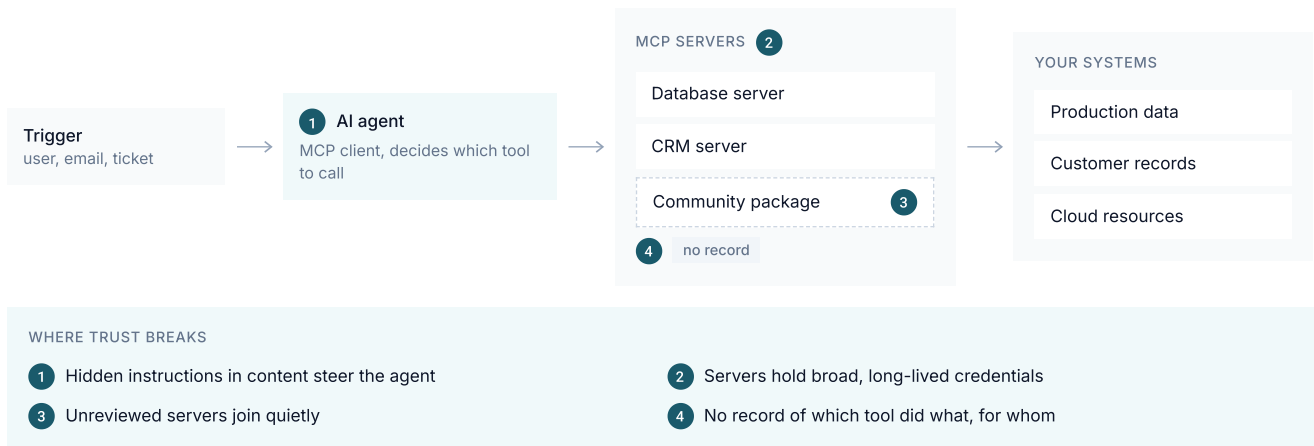


Figure 1 · How an MCP deployment works, and where trust breaks

Three ways it goes wrong

- **The steered agent.** A coding agent reads a support ticket containing hidden instructions, and runs a destructive command against a shared environment.
- **The over-scoped server.** A CRM server built for reading accounts also holds write and delete permissions, so one bad tool call can change customer records.
- **The shadow server.** A developer installs a community MCP package that nobody reviewed. Researchers have already documented malicious MCP packages published to public registries.

Why IAM and API security are not enough

Identity systems authorize the agent once, when it connects. API gateways see requests but not the intent behind a tool call or the arguments inside it. Neither can answer the question that matters for agents: should this specific action, by this agent, with these arguments, happen right now? That question needs an answer on every single call.

3. The MCP threat landscape

MCP risk enters at three points: before a server ever runs, in who is allowed to do what, and inside each tool call.

The OWASP MCP Top 10 names these risks for the protocol itself. It is still in beta, so we pair it with the finalized OWASP Top 10 for Agentic Applications, which covers agent behavior more broadly, including agent goal hijack (ASIO1), agentic supply chain vulnerabilities (ASIO4), unexpected code execution (ASIO5), and rogue agents (ASIO10). The Coalition for Secure AI's MCP security paper adds a detailed threat taxonomy and secure design patterns.

The figure below groups the OWASP MCP risks by where they enter, with badges showing which of the eight controls in chapter 4 reduce each one. One risk, missing audit and telemetry, cuts across everything: without a record of tool calls, every other risk becomes harder to detect and impossible to prove.

FIGURE 02
MCP risk enters at three points, and missing logs hide all of them

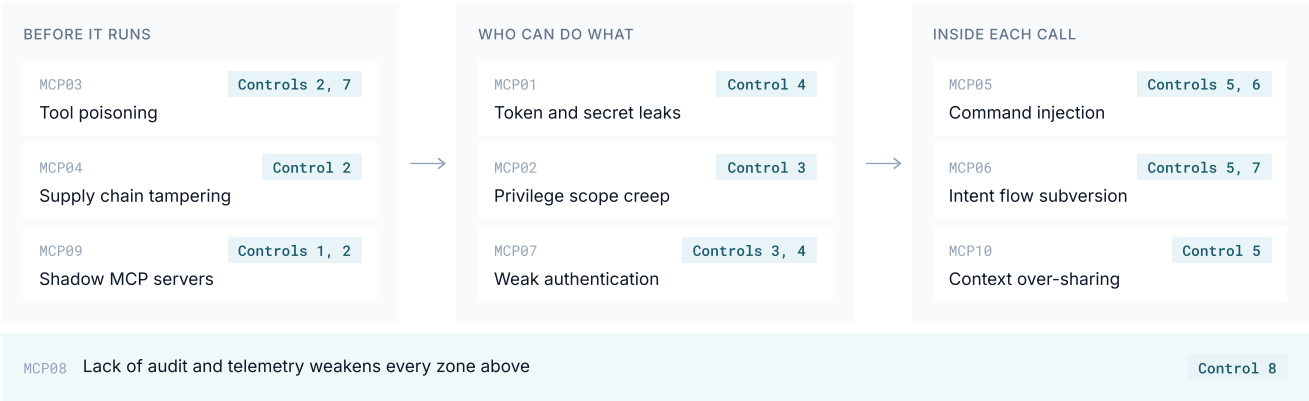


Figure 2 · OWASP MCP Top 10, grouped by where each risk enters

4. The 8 controls for MCP runtime governance

Eight controls, applied at one checkpoint between agents and every MCP server, turn invisible agent actions into governed, recorded decisions.

They work as a set. The first two make sure you know every server and trust what it offers. The next four decide, on every call, whether an action may happen and with what. The last two watch behavior over time and prove what happened.

FIGURE 03
Eight controls: two discover, four enforce, two watch and prove

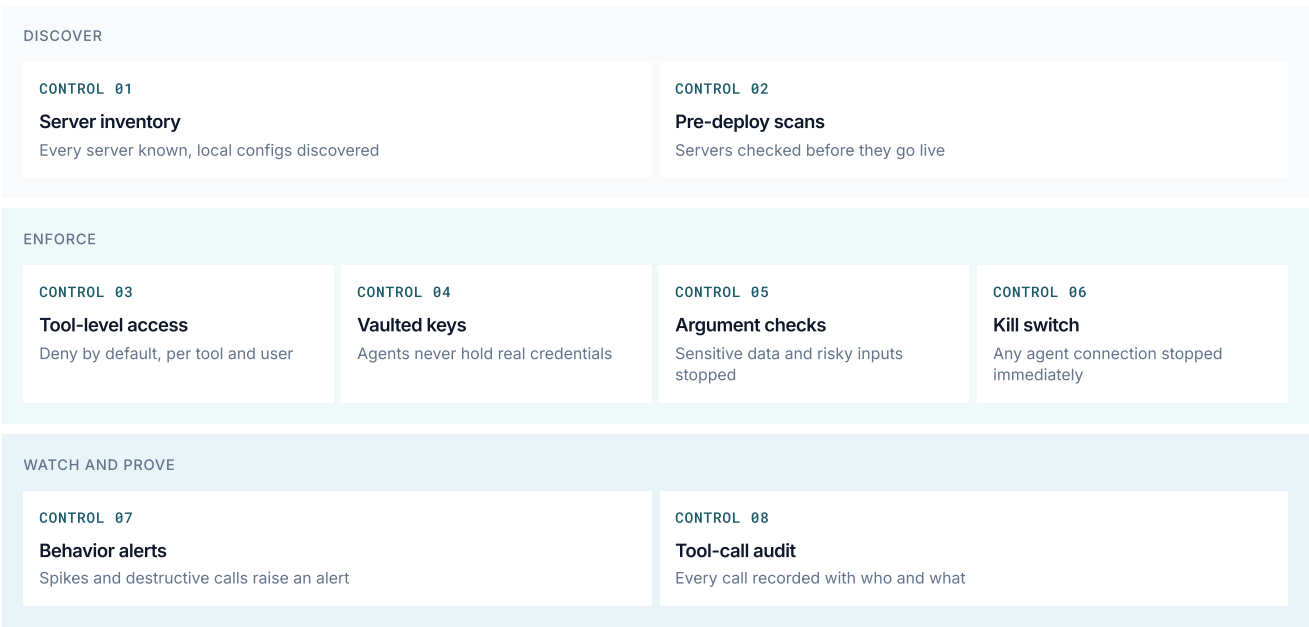


Figure 3 · The eight MCP controls at a glance

FIGURE 04
Every tool call is checked, then given a key, then recorded

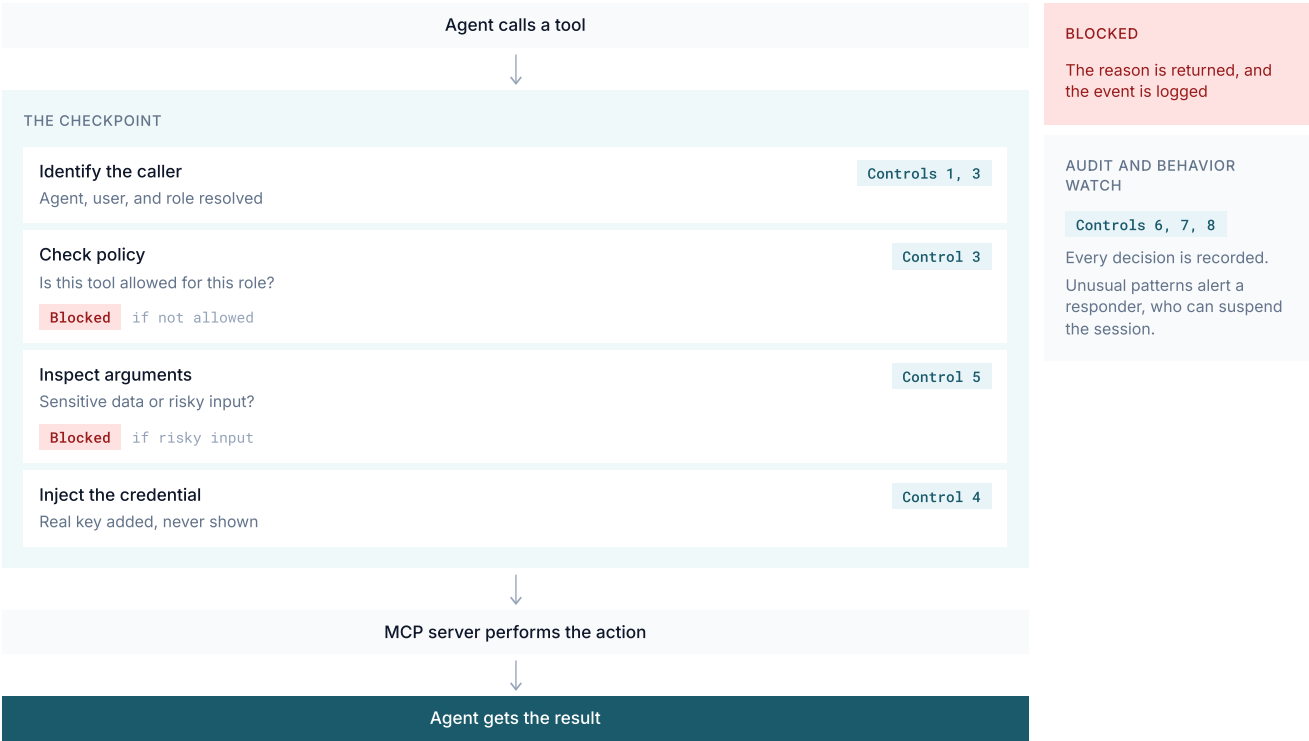


Figure 4 · The life of one tool call through the checkpoint

The agent never holds the real key, a blocked call always returns a reason, and every decision lands in the audit trail.

5. Framework crosswalk

Auditors rarely ask about MCP by name yet, but every framework they use already asks the questions MCP raises: who can act, with what access, and where is the record?

The table maps each control to the clauses it supports. NIST is also developing control overlays and identity guidance specifically for AI agents, so expect agent-level expectations to sharpen. Building these controls now puts you ahead of that curve. As always, no tool makes you compliant on its own; the right layer supplies the controls and the evidence.

CONTROL	STANDARDS (NIST, ISO, ETSI)	EU (AI ACT, GDPR, DORA)	INDIA (DPDP, AI GUIDELINES, CERT-IN)	INDUSTRY (SOC 2, HIPAA, PCI DSS, GLBA)
1. Discovery and inventory	NIST AI RMF Govern (AI inventory); ISO/IEC 42001 operational planning; ISO/IEC 27001 A.5.9 asset inventory	AI Act Art. 26 deployer duties; DORA register of ICT third-party providers	DPDP accountability for processors; Accountability sutra	SOC 2 CC6.1 asset inventory
2. Pre-deployment scanning	NIST Cyber AI Profile Identify; ISO/IEC 27001 A.8.8 technical vulnerabilities and A.5.21 ICT supply chain; ETSI EN 304 223	AI Act Art. 15 cybersecurity; DORA ICT third-party risk	Safety, Resilience and Sustainability sutra	SOC 2 CC7.1 vulnerability detection; PCI DSS Req. 6 and 11
3. Tool-level access control	NIST AI RMF Manage; ISO/IEC 27001 A.5.15 access control and A.8.2 privileged access	GDPR Art. 32 security of processing; AI Act Art. 26 use as intended	DPDP reasonable security safeguards	SOC 2 CC6.1 and CC6.3; HIPAA access control 164.312(a); PCI DSS Req. 7

CONTROL	STANDARDS (NIST, ISO, ETSI)	EU (AI ACT, GDPR, DORA)	INDIA (DPDP, AI GUIDELINES, CERT-IN)	INDUSTRY (SOC 2, HIPAA, PCI DSS, GLBA)
4. Credential injection	NIST Cyber AI Profile Protect; ISO/IEC 27001 A.5.17 authentication information and A.8.24 cryptography	GDPR Art. 32	DPDP reasonable security safeguards	SOC 2 CC6.1; PCI DSS Req. 8 authentication
5. Argument and data checks	NIST AI RMF Measure (privacy); ISO/IEC 27001 A.8.12 data leakage prevention; ISO/IEC 27701	GDPR Art. 5(1)(c) minimization and Art. 25 privacy by design	DPDP purpose limitation and safeguards	HIPAA minimum necessary; PCI DSS Req. 3; GLBA safeguards
6. Session control and kill switch	NIST Cyber AI Profile Respond	AI Act Art. 14 human oversight; DORA incident management	People First sutra (human oversight)	SOC 2 CC7.4 incident response; HIPAA security incident procedures
7. Behavioral anomaly detection	NIST Cyber AI Profile Detect; ISO/IEC 27001 A.8.16 monitoring	GDPR Art. 33 breach notification; DORA incident management	CERT-In incident reporting; DPDP breach intimation	SOC 2 CC7.2 and CC7.3
8. Tool-call audit trail	NIST AI RMF Govern and Measure; ISO/IEC 42001 Clause 9; ISO/IEC 27001 A.8.15 logging	AI Act Art. 12 record keeping and Art. 26 log retention; GDPR Art. 30 records of processing	CERT-In log retention; Accountability sutra	SOC 2 CC7.2; HIPAA audit controls 164.312(b); PCI DSS Req. 10

A note on the EU AI Act. Articles 12, 14, and 15 apply in law to high-risk AI systems. For every other agent deployment, they remain the clearest public benchmark for logging, human oversight, and security, and enterprise buyers increasingly read them that way.

6. The MCP Governance Maturity Model

Most teams adopting agents start at level one, where nobody can list every MCP server in use, and each level up closes a door attackers rely on.

FIGURE 05
Each maturity level closes a door attackers rely on

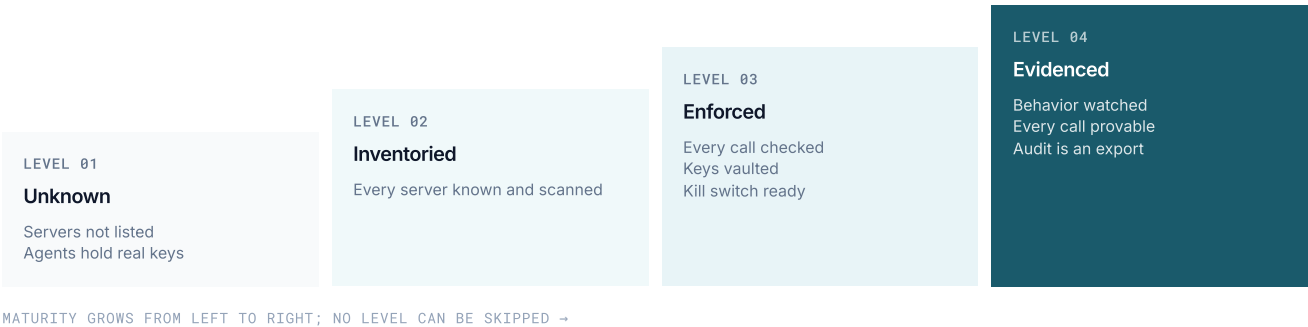


Figure 5 · The MCP Governance Maturity Model

Score yourself

Tick each statement that is true for your organization today.

- ☐ We know every MCP server our agents can reach, including ones developers installed.
- ☐ Every server is scanned before it goes live.

- ☐ Agents never hold the real credentials for the systems they use.
- ☐ Every tool call is checked against a deny-by-default policy.
- ☐ A responder can stop any agent session immediately.
- ☐ Every tool call is logged with user, agent, tool, and arguments, and can be exported.

How to read your score. Zero or one tick: Level 1, Unknown. Two or three: Level 2, Inventoried. Four or five: Level 3, Enforced. All six: Level 4, Evidenced.

7. Industry playbooks

Every industry runs different agents, so the first controls to switch on differ too.

INDUSTRY	AGENTS IN USE	THE RISK THAT MATTERS MOST	START WITH
Fintech and banking tech	Reconciliation and reporting agents connected to bank APIs, ledgers, and ERP	An agent moving money or changing financial records without approval	Controls 3, 6, then 8
Healthtech	Agents connected to scheduling, patient records, and billing	Patient records read or shared beyond the task at hand	Controls 5, 3, then 8
Legal tech	Research and drafting agents connected to document stores and matter systems	Privileged documents pulled into the wrong matter or sent outside	Controls 3, 5, then 8
Enterprise SaaS	DevOps, support, and CRM agents connected to code, tickets, and customer data	Shadow servers and over-scoped credentials spreading across engineering	Controls 1, 2, then 4

8. A three-phase rollout

Govern agents the way you would onboard a new employee: first learn what they touch, then set the rules, then keep the records.

FIGURE 06
Discover, enforce, prove: each phase stands on its own

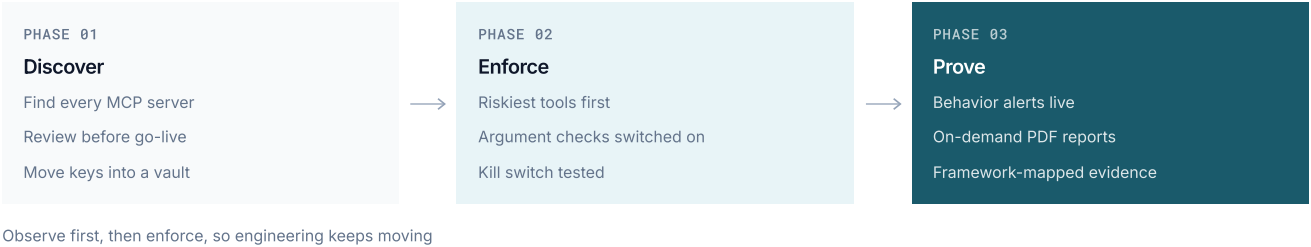


Figure 6 · The three-phase rollout

Start with log-and-allow (alert) rules, so engineering sees no friction while security gets its first complete picture. Then switch on policies for the riskiest tools first, such as anything that deletes, writes, or moves money.

9. Buyer's checklist

Before choosing any MCP security platform, ask these eight questions; a strong vendor answers each one with a demo, not a slide.

1. Can it show every MCP server our agents use, including ones developers installed themselves?
2. Does it scan servers and tool definitions before they go live?
3. Is every tool call deny-by-default, with policy at the level of individual tools?
4. Are real credentials injected at the gateway, so agents never hold them?
5. Can policies inspect tool arguments, not just tool names?
6. Can a responder suspend an agent session immediately?
7. Does it flag unusual behavior, such as a never-before-seen tool or a sudden spike?
8. Is every call logged and exportable as evidence, and can we retain it for as long as our auditors need?

10. How Igris delivers this

Igris puts seven of the eight controls between your agents and every MCP server, and developers adopt it with a TypeScript SDK or by pointing their MCP client at Sentinel's URL with an Igris API key, rather than a rebuild.

Two Igris products carry this paper's controls. **Igris Sentinel** is the MCP gateway that every tool call passes through: it identifies the caller, matches policy, injects credentials, routes the call, and records it. **Igris Lens** turns every event into one timeline with incidents, alerts, and audit-ready reports.

FIGURE 07

Igris sits between every agent and every MCP server, and records it all



Figure 7 · Where Igris Sentinel and Lens sit in your agent stack

Every control, mapped

CONTROL	HOW IGRIS DELIVERS IT
1. Discovery and inventory	Sentinel becomes the single path to your MCP servers, so every connection, agent, and tool routed through it appears in Lens; the Igris CLI finds MCP configs on developer machines to bring them under governance
2. Pre-deployment scanning	Not automated by Igris today. Because agents can only reach servers registered with Sentinel, registration becomes the review gate: pair it with your existing security review before a server goes live.
3. Tool-level access control	Deny-by-default, first-match policies with glob patterns on tool names and conditions on user and request metadata such as role
4. Credential injection	An encrypted credential store (AES-256-GCM) holds upstream keys, and Sentinel injects them on the wire, so real tokens never reach the agent or the audit log
5. Argument and data checks	Content Guard scans tool arguments for sensitive data patterns and blocked keywords, and redacts them before the call reaches the server
6. Session control and kill switch	Suspend any agent connection in one click; anomaly alerts tell responders when to act
7. Behavioral anomaly detection	Rate-spike and destructive-action detection flags unusual agent behavior, with alerts to Slack, Discord, or any HTTP webhook
8. Tool-call audit trail	Every call recorded with user, tool, arguments, and decision in a retained audit trail, exportable as CSV, with JSON evidence through the SOC 2 export

Sentinel ships a TypeScript SDK, and any language can reach it over HTTP.

See it on your own agents

The best way to judge any of this is on your own agent traffic. Book a demo at igrisecurity.com and we will walk through the eight controls live, starting with Phase 1: discover every server.

Sources

- **Agent and MCP security:** [OWASP MCP Top 10 \(beta\)](#), [OWASP Agentic Security Initiative](#), [CoSAI MCP Security guide](#), [MITRE ATLAS](#)
- **NIST:** [AI Risk Management Framework](#), [AI 600-1](#), [IR 8596 Cyber AI Profile \(preliminary draft\)](#), [AI Agent Standards Initiative](#), [COSAiS control overlays \(in development\)](#)
- **ISO and ETSI:** [ISO/IEC 42001](#), [ISO/IEC 27001](#), [ISO/IEC 27701](#), [ETSI EN 304 223](#)
- **European Union:** [EU AI Act](#), [GDPR](#), [DORA](#)
- **India:** [DPDP Rules \(obligations phasing in through 2027\)](#), [India AI Governance Guidelines](#), [CERT-In Directions](#)
- **Industry and assurance:** [SOC 2 Trust Services Criteria](#), [HIPAA Security Rule](#), [PCI DSS](#), [GLBA Safeguards](#)