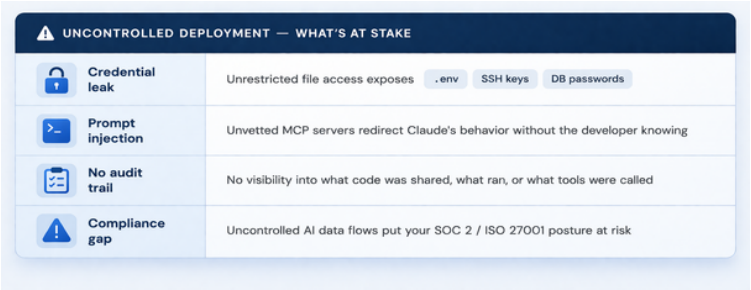
truefoundry

Enterprise Security for *Claude Code*

CISO & Security Admin Guide
CLI · Models · Tools · MCP · Audit

Your developers are already using it. Have your security controls caught up?

Claude Code runs with the same OS permissions as the local user: reading files, executing shell commands, connecting to external MCP servers, and sending code to Anthropic's servers. None of this surfaces in your existing DLP or endpoint controls. And with SOC 2, ISO 27001, and emerging AI governance frameworks now asking how AI tools access and transmit data, the compliance clock is already ticking.



WHAT YOU NEED TO CONTROL — AND HOW TRUEFOUNDRY MAKES IT ENFORCEABLE

1 Identity & API Key Governance

Invisible usage and unrevoked access are the first failure mode

Developers authenticating with personal Anthropic accounts have no corporate oversight. Usage is invisible, keys are unmanaged, and offboarding leaves access open indefinitely. Enforce SSO **(SAML 2.0 / OIDC)** via the Claude Admin Console with domain capture. Manage all API keys centrally by storing in Vault or AWS Secrets Manager, rotating quarterly and revoking on offboarding.

2 Model Access & Traffic Routing

Without a proxy, all Claude traffic is a blind spot

Developers can freely switch to unapproved models, and all traffic goes directly to Anthropic with no organizational visibility or control point. Route all Claude Code traffic through a centralized gateway and lock model selection via MDM-deployed **managed-settings.json** — tamper-resistant at the OS level.

Enforcing Model Routing Across Developer Machines

Setting environment variables on one machine is easy. The real challenge is enforcing them consistently across all developers. There are three approaches, each with its own trade-offs.

- MDM/Endpoint managed
- Server-managed settings (beta)
- Direct cloud provider (Bedrock/Vertex)

Approach	Enforcement	Security Level
MDM + TrueFoundry AI Gateway	OS-level settings lock + centralized gateway	✅ Very High
MDM + Anthropic Direct	OS-level settings lock + direct API	🟧 High
MDM + AWS Bedrock	Cloud IAM only	🟧 High
MDM + Google Vertex AI	OS-level settings lock + cloud IAM	✅ Very High
Server-Managed Settings	Admin Console (client-side)	⚠️ Medium
Direct AWS Bedrock	Cloud IAM only	🟧 High
Direct Google Vertex AI	Cloud IAM only	🟧 High
TrueFoundry AI Gateway (no MDM)	Gateway-level controls	🟧 High

3 Configure Tool Access & Sandboxing

By default, Claude Code can do anything the developer can

Claude Code runs directly in a developer's terminal with full user-level permissions, accessing files, executing shell commands, and reaching external services, giving it a very large attack surface.

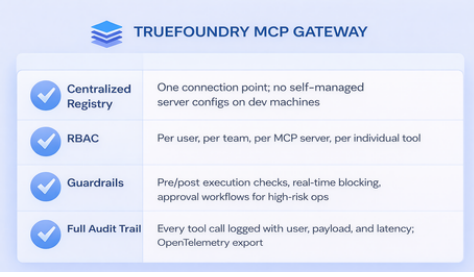
Sandboxing

Set **"allowUnsandboxedCommands":** false in sandbox settings to disable the escape hatch that allows commands to run outside the sandbox. This forces all commands to run sandboxed or be explicitly listed in **excludedCommands**.

4 MCP Server Governance

Shadow IT and prompt injection enter through unvetted tool connections

MCP servers extend Claude Code's reach to databases, internal tools, and external APIs. Without oversight, developers self-install unvetted public servers, credentials sprawl across machines, and there's no record of what data was accessed or exfiltrated. Route all MCP traffic through a centralized gateway and allowlist only the gateway URL in managed settings.



Set **strictKnownMarketplaces** to an empty array ([]) to block all marketplace-sourced MCP installations unless you have explicitly reviewed and approved them.

Ready to govern Claude Code across your org?

TrueFoundry's AI Gateway and MCP Gateway give you a single control plane — identity, model access, tool governance, and audit logging — without blocking developer productivity.

Full config docs: truefoundry.com/docs/ai-gateway/mcp/enterprise-security-claude

Scan to book a demo

truefoundry

Enterprise Security for *Claude Code*

CISO & Security Admin Guide
CLI · Models · Tools · MCP · Audit

Enforcing Model Routing Across Developer Machines

Without enforcement, developers can connect Claude Code to any MCP server they choose, including unvetted public servers. These two approaches close that gap:

- **Apply via MDM (recommended):** Deploy **managed-mcp.json** to every managed device. Developers cannot add or connect to any MCP server beyond what's defined. **allowManagedPermissionRulesOnly: true**, ensures system-level rules become the only rules. Project and user settings cannot override them.
- **Apply via Server-Side setting:** For BYOD or unmanaged devices, configure MCP allowlists directly in the Claude Admin Console. Settings apply to all org members on authentication — no file deployment needed.

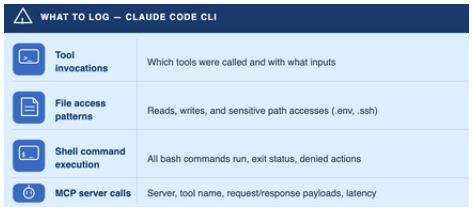
5 Data Retention

Source code sent to an LLM is a data flow your auditors will ask about

By default, Anthropic may retain prompts and outputs. Local session transcripts persist indefinitely on developer machines — a data residency and IP exposure risk. Review your enterprise agreement for Anthropic's retention scope. Set **transcriptRetentionDays** in managed settings to auto-delete local transcripts (recommended: 7–14 days).

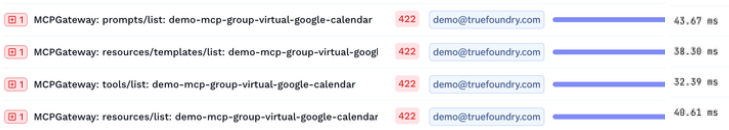
6 Configure audit logs and monitoring

If you can't answer what ran and what was shared, you can't pass an audit



Without centralized logging, you have no answer when auditors ask: who used Claude Code, what code was shared, which tools were called, what commands executed. Route traffic through TrueFoundry's gateway for complete session attribution — tool invocations, file access, shell commands, denied actions, and MCP calls.

Export logs to your SIEM on a regular schedule. Retain for a minimum of 90 days to meet SOC 2 requirements; longer for regulated industries.



7 Configure Custom Usage

Set hard monthly spending limits per user and per team to prevent runaway costs.

Without usage controls, AI adoption creates both financial exposure and blind spots in accountability. You need enforceable limits and clear visibility into who is consuming resources—and how.

Usage Limits
Set hard monthly spending limits per user and team to prevent runaway costs and uncontrolled scaling.

Usage Visibility
Maintain centralized reporting across users, teams, and models to:

- Identify high-cost usage for governance and optimization
- Attribute spend for chargeback or showback
- Detect anomalous spikes that may indicate misuse or compromise

8 Compliance and Regulatory Controls

Claude Code introduces new data flows your auditors will want to trace

Claude Code operates with local OS access, executes commands, and transmits code externally—making it a regulated data path under SOC 2, HIPAA, and emerging AI governance frameworks. You must explicitly control and document how it is used.

SOC 2 Type II
Anthropic is SOC 2 Type II certified, but Claude Code usage sits within your control boundary. Ensure controls are enforceable and auditable:

- Enforce provisioning and deprovisioning for all developer access
- Retain and export audit logs (90+ days) covering code access, commands, and tool usage
- Maintain ongoing vendor risk assessments and contractual oversight

HIPAA
Claude Code can access local files and transmit sensitive data, making PHI exposure a real risk. Do not allow PHI usage without a signed Zero Data Retention (ZDR) addendum.

- Require human review for any outputs involving patient data
- Maintain a complete audit trail of all PHI-related interactions
- Document Claude Code usage in your HIPAA risk analysis

Treat every interaction as a potential cross-border data transfer subject to GDPR controls and accountability.

Requirement	Implementation
Data residency	Deploy via AWS EU regions or Google Vertex AI with Private Service Connect
Right to erasure	Use the Compliance API and documented deletion workflows; coordinate DSARs with Anthropic
Data minimization	Configure `deny` rules (CLI) and upload restrictions (Web) to block PII-containing content
Purpose limitation	Document approved AI processing use cases in your Records of Processing Activities
Purpose limitation	Document approved AI processing use cases in your Records of Processing Activities

Ready to govern Claude Code across your org?

TrueFoundry's AI Gateway and MCP Gateway give you a single control plane — identity, model access, tool governance, and audit logging — without blocking developer productivity.

Full config docs: truefoundry.com/docs/ai-gateway/mcp/enterprise-security-claude

Scan to book a demo