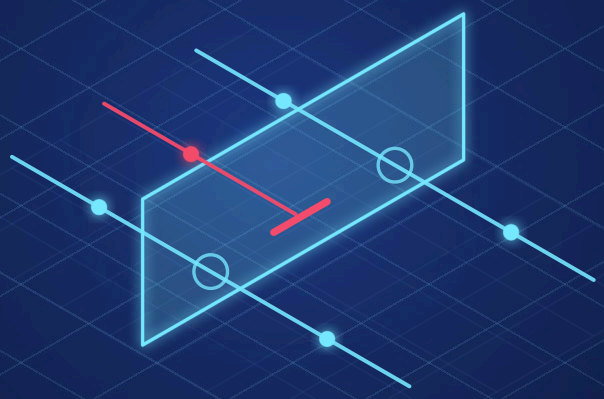
SecureAuth

AGENT AUTHORITY

Securing AI agent access against the OWASP Agentic Top 10

A guide for security and IAM leaders on keeping AI agents accountable.



© 2026 SecureAuth Corporation.

Ebook

CONTENTS

Inside this ebook

01 **Agent access is a delegation problem** 3
Why agents need their own identity, and how delegation and tool class set the level of control.

02 **Reference architecture** 5
Where Agent Authority sits, and how a tool call is checked.

03 **OWASP risks at a glance** 7
All ten OWASP agentic risks on one page, by path and control.

04 **How each risk is addressed** 8
For each risk: how it shows up, the controls that apply and what stays outside them.

05 **Safe defaults** 10
One tool call traced through each point where policy decides.

06 **Control catalog** 11
Controls AA-01 to AA-14 with mechanism, default outcome and evidence.

07 **Framework alignment** 12
An indicative mapping to NIST SP 800-53 control families.

AUDIENCE
Security, IAM and technology risk leaders evaluating agent access controls.

COVERS
AI agents on employee devices, their tool calls over MCP and their model traffic.

CONVENTIONS
Controls are AA-01 to AA-14, and risks use OWASP ASI IDs.

© 2026 SecureAuth Corporation.

CHAPTER 1

Agent access is a delegation problem

AI agents act with the authority of the people who run them. That delegation needs its own identity and a decision on each request.

Coding agents, desktop assistants and internal agents now open pull requests, query data stores, file tickets and post messages. Each action uses an employee's authority, and most leave the device through one of two channels: a tool call over the Model Context Protocol, or a request to a model provider. Some stay on the device, such as a local MCP server or a shell command run by a coding agent.

The access patterns that grew up around these agents favor speed. Personal access tokens sit in configuration files, shared service accounts stand in for teams, and local servers are installed without review. Access then resolves to a token, with no record of the person an agent acted for or the policy that allowed the action.

Existing access controls miss two things about agents. Each action has two actors, the person and the client acting for that person. The intent behind each action is generated at run time by a model, and the content the model reads can steer it. Agent Authority binds the client to the person, decides each request with its arguments in view and records both actors with the decision.

- **Bound identity**
  Each client is registered to the person who authorized it, or flagged as acting alone.
- **Per-request decision**
  Each call is decided with its tool, arguments and network context in view.
- **Safe default**
  A rule or filter that cannot evaluate refuses the request.
- **Record of both actors**
  Each decision names the person, the instance and the deciding rule.

KEY TERM

An *agent instance* is one connected client of an agent product, bound to the person who authorized it.

WHY DECIDE PER REQUEST

Consent happens once, at sign-in. An agent chooses a new action at each step, so authorization has to see each request and its arguments.

© 2026 SecureAuth Corporation.

Delegation and tool class set the level of control

Two attributes decide how much control an agent action needs, and both can be read from data an IAM team already holds. The first is whether a person authorized the agent instance, which Agent Authority records as a read-only autonomous flag. The second is what the action can change, which tool classification expresses as read-only, write, destructive or external. Together they form a grid that maps onto enforceable rules, shown in Figure 1.

STARTING POSTURE

Each cell becomes one ordered rule. Teams then add argument and network conditions.

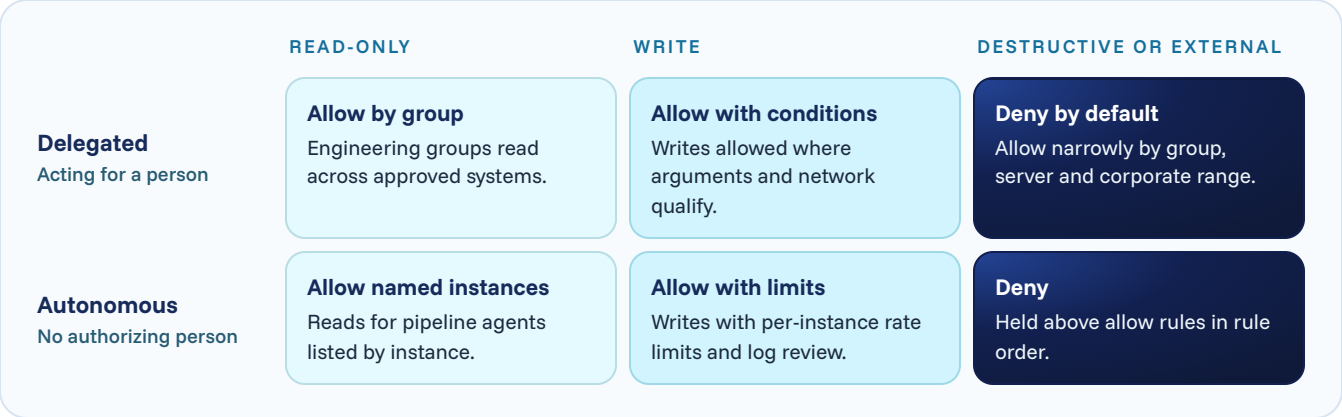


Figure 1. A starting posture by delegation and tool class.

One identity across the device, tool and model paths

Agent activity reaches the enterprise through three paths, and each has its own enforcement point in Agent Authority. Endpoint Authority runs on the device and checks actions that stay on the device. MCP Authority brokers tool calls to internal and SaaS systems. LLM Authority sits in front of model providers. Each path reads the same person, instance and group context and writes to the same Agent Trail, so a review of one employee's agents covers each path.

ASSUMPTIONS

The enterprise identity provider is authoritative for people and groups. Upstream systems keep enforcing their own entitlements. Model traffic from configured AI tools passes through LLM Authority.

Device path

ENDPOINT AUTHORITY

Local MCP servers, shell commands and outbound destinations, checked before they run.

Tool path

MCP AUTHORITY

Tool calls to internal and SaaS systems, decided per request on one brokered endpoint.

Model path

LLM AUTHORITY

Prompts to model providers, redacted and checked for injection, within team budgets.

Chapter 2 shows the architecture. Chapter 3 lists the OWASP risks on one page, and chapter 4 shows how each one is addressed. Chapters 5 and 6 cover safe defaults and the control catalog, and chapter 7 maps the controls to NIST SP 800-53.

© 2026 SecureAuth Corporation.

Agent Authority applies one policy on each path an agent can take

Agent Authority has an enforcement point on the device, on tool calls over MCP and on prompts to model providers. Each point reads the same identity and policy, and each decision lands in one Agent Trail.



Figure 2. Agent Authority reference architecture. Checks at each enforcement point and the evidence they leave are highlighted.

© 2026 SecureAuth Corporation.

MCP Authority in depth: each control on the path to an enterprise system

Each tool call passes the gateway, where policy decides it, the employee's credential is applied, the response is filtered and the decision is recorded.

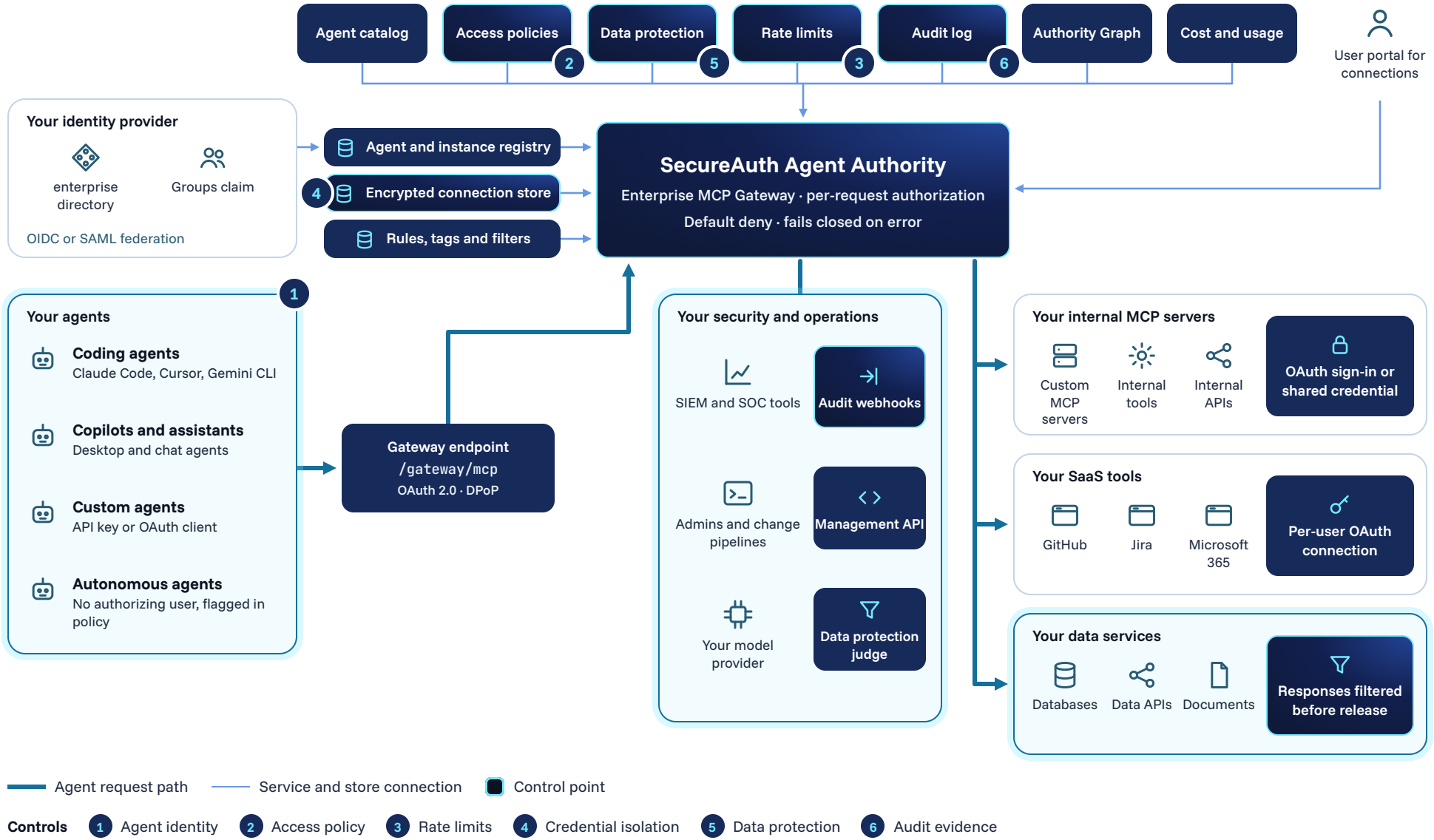


Figure 3. Agent Authority reference architecture, with the control points a risk team reviews outlined and numbered.

© 2026 SecureAuth Corporation.

CHAPTER 3

OWASP risks at a glance

All ten risks from the OWASP Top 10 for Agentic Applications, with the paths where each appears and the controls that apply. Chapter 4 covers each risk in detail.

ID	RISK	HOW IT APPEARS IN AGENT ACCESS	DEVICE	TOOL	MODEL	CONTROLS
ASI01	**Agent goal hijack**	Instructions planted in a ticket, document or tool result redirect the agent toward a harmful action.	○	●	●	AA-09, AA-03, AA-04, AA-05
ASI02	**Tool misuse and exploitation**	An allowed tool is called with harmful arguments or at harmful volume.	●	●	○	AA-04, AA-05, AA-10, AA-12
ASI03	**Identity and privilege abuse**	An agent inherits a broad token or shared account and acts beyond the person's entitlements.	○	●	●	AA-01, AA-02, AA-03, AA-06
ASI04	**Agentic supply chain vulnerabilities**	A tool definition changes after approval, or an unreviewed local MCP server is installed.	●	●	○	AA-08, AA-12, AA-03
ASI05	**Unexpected code execution**	A coding agent runs shell commands or destructive repository operations on a laptop.	●	○	○	AA-12, AA-14
ASI06	**Memory and context poisoning**	Sensitive or adversarial content enters the agent's context through results or prompts.	○	●	●	AA-07, AA-09
ASI07	**Insecure inter-agent communication**	One agent invokes another through a channel that neither authenticates nor attributes.	○	●	○	AA-01, AA-03, AA-14
ASI08	**Cascading failures**	A loop or misread error multiplies calls and spend across systems in minutes.	○	●	●	AA-10, AA-11, AA-13
ASI09	**Human-agent trust exploitation**	A person accepts agent output without seeing the actions taken to produce it.	●	●	●	AA-14
ASI10	**Rogue agents**	A compromised, unapproved or misconfigured agent acts while appearing legitimate.	●	●	●	AA-12, AA-11, AA-01, AA-02

● Risk appears on this path ○ Not a primary path for this risk Control IDs refer to the catalog in chapter 6.

Boundaries of the model. Assets are enterprise systems reached by agents and the data they return. Actors are employees, their agents, autonomous agents and content that can steer a model. Trust boundaries sit at the device, the gateway and the model endpoint.

© 2026 SecureAuth Corporation.

04 CHAPTER 4

How each risk is addressed

One row per OWASP risk: how it shows up, the controls that apply and what stays outside them.

ASI01

Agent goal hijack

DEVICE TOOL MODEL

Instructions hidden in a ticket or document redirect the agent.

AA-09	Prompt injection detection	AA-03	Each call checked against policy
AA-05	Outbound tools blocked by class		

OUTSIDE SCOPE Detection is probabilistic. Rules bound the outcome.

ASI02

Tool misuse and exploitation

DEVICE TOOL MODEL

An approved tool is used to delete records or overwrite code.

AA-04	Rules on what a call contains	AA-05	Destructive tools classified
AA-12	Laptop actions checked first	AA-10	Call limits

OUTSIDE SCOPE Upstream branch protection stays in place.

ASI03

Identity and privilege abuse

DEVICE TOOL MODEL

An agent runs on a shared account or a long-lived personal token.

AA-01	Each agent tied to a person	AA-02	Copied tokens refused
AA-06	Systems see the real employee	AA-03	Deny unless allowed

OUTSIDE SCOPE Shared-key servers attribute per resource upstream.

ASI04

Agentic supply chain

DEVICE TOOL MODEL

A tool changes after approval, or an unreviewed local server appears.

AA-08	Changed tools held for re-approval	AA-12	Unapproved local tools blocked

OUTSIDE SCOPE Servers enter the catalog through review.

ASI05

Unexpected code execution

DEVICE TOOL MODEL

A coding agent runs commands on a laptop that change code or systems.

AA-12	Laptop actions checked first	AA-14	Recorded in the Agent Trail

OUTSIDE SCOPE Host malware protection stays with endpoint tools.

© 2026 SecureAuth Corporation.

ASI06

Memory and context poisoning

DEVICE TOOL MODEL

Sensitive or adversarial content enters the agent's context.

AA-07 **Sensitive data removed from responses**

AA-09 **Prompts redacted before they leave**

OUTSIDE SCOPE Memory held inside the agent is not inspected.

ASI07

Insecure inter-agent communication

DEVICE TOOL MODEL

One agent calls another with no authentication or attribution.

AA-01 **Each agent authenticates**

AA-03 **Each call checked**

AA-14 **Both actors recorded**

OUTSIDE SCOPE Channels outside MCP need their own authentication.

ASI08

Cascading failures

DEVICE TOOL MODEL

A loop multiplies calls and spend within minutes.

AA-10 **Call limits**

AA-11 **Unusual activity flagged**

AA-13 **Spend budgets**

OUTSIDE SCOPE Limits are sized per environment, starting in monitor-only mode.

ASI09

Human-agent trust exploitation

DEVICE TOOL MODEL

A person accepts agent output without seeing what the agent did.

AA-14 **Actions and decisions shown per person**

OUTSIDE SCOPE Output accuracy sits outside access control.

ASI10

Rogue agents

DEVICE TOOL MODEL

A compromised or unapproved agent acts as if it were legitimate.

AA-12 **Shadow AI discovery**

AA-11 **Unusual activity flagged**

AA-01 **Access revoked at the next request**

AA-14 **SIEM correlation**

OUTSIDE SCOPE Device discovery finds agents that bypass the gateway.

WHAT THIS MEANS FOR YOUR PROGRAM

Access reviews

Each agent maps to a person and a directory group, so agents appear in the reviews you already run.

Audit

Allowed and refused calls carry the person, agent and rule, ready for the mapping in chapter 7.

Adoption

Agents connect through configuration with no code change, so approved agents reach teams sooner.

Cost

Spend is attributed to teams and people, with budgets that stop requests at the limit.

© 2026 SecureAuth Corporation.

One MCP tool call, traced through each decision point

The MCP path carries the most decisions, so it is traced here end to end. The device and model paths read the same identity and policy and write to the same Agent Trail.

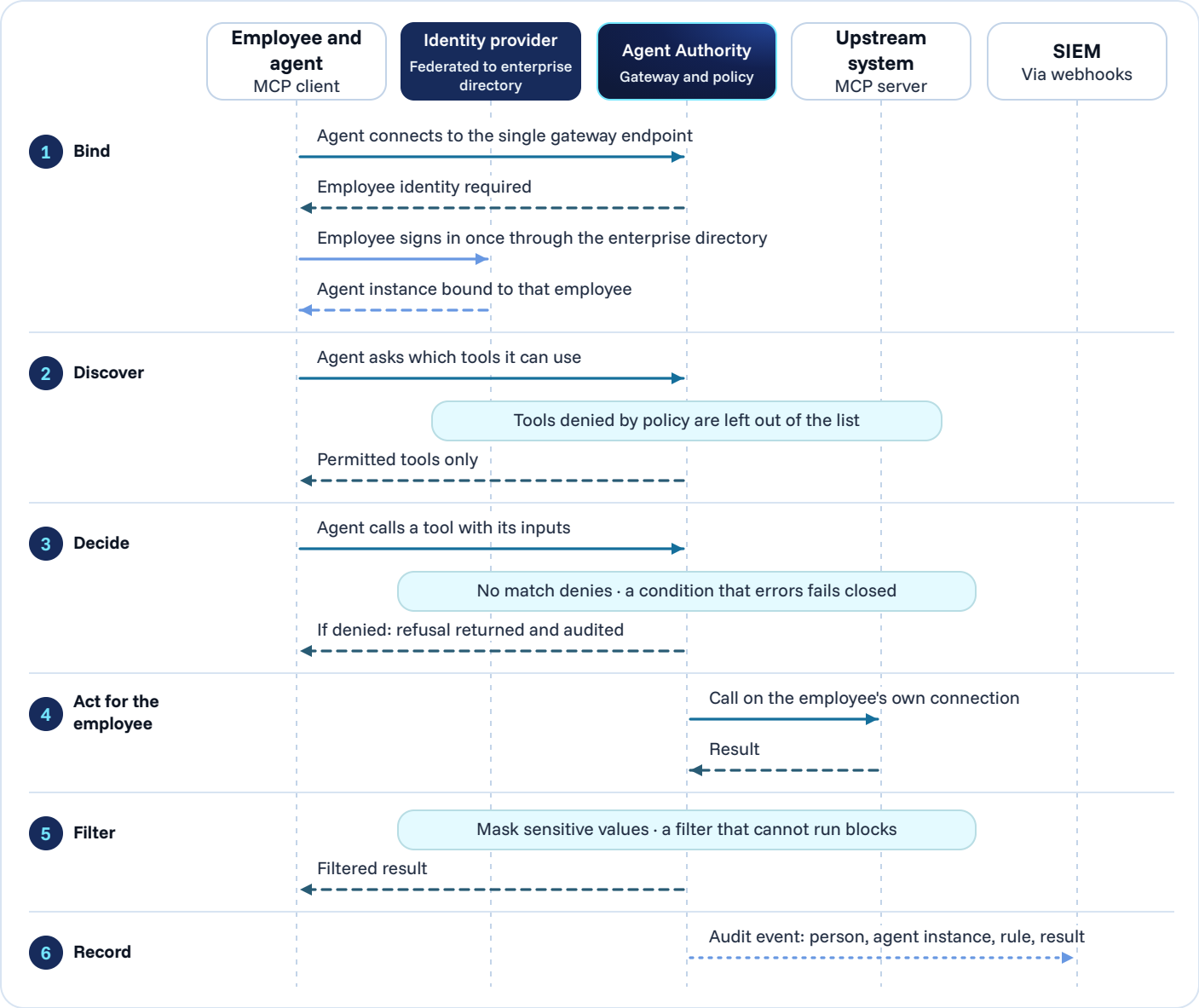


Figure 4. Request sequence for an agent acting for an employee. Dashed arrows are responses.

POINTS FOR SECURITY REVIEW

3 Default deny, fail closed
No matching rule means deny. A condition that errors skips allow rules and keeps deny rules in force.

4 Credentials out of reach
Upstream credentials are encrypted, isolated per organization and kept out of the agent's reach.

5 Filters block by default
A filter that cannot run blocks the response by default, as does a response too large to inspect.

6 Records without raw content
Gateway audit events hold metadata, and redactions are recorded as rule names and counts with the matched text left out.

© 2026 SecureAuth Corporation.

CHAPTER 6

Control catalog

Each control lists how it works, its default outcome when a check cannot complete and the evidence it leaves for review.

ID	CONTROL	HOW IT WORKS	DEFAULT OUTCOME	EVIDENCE
AA-01	**Instance binding**	Each agent client is registered to the employee who authorized it. Agents acting with no person behind them are flagged.	**Revoked or unbound tokens are refused at the next request.**	Connection and revocation events naming person and instance
AA-02	**Client authentication**	Agents sign in with standard OAuth through your identity provider. Tokens can be bound to the client so a copied token cannot be reused.	**A copied, unbound or unauthorized token is refused.**	Refused authentication attempts in the audit log
AA-03	**Ordered rules**	Rules are read top to bottom and the first match decides. Deny rules sit above allow rules.	**No matching rule: the call is denied.**	Deciding rule on each call, with a default-deny group
AA-04	**Conditions**	Rules can check the person's groups, what the call contains and the network it comes from.	**If a rule check errors, the allow does not apply and deny rules hold.**	Rule and effect on each call
AA-05	**Tool classification**	Tools are classed as read-only, write, destructive or external, so one rule covers a whole class.	**Unclassified tools stay denied unless another rule allows them.**	Effective access per agent type
AA-06	**Brokered credentials**	Agents do not hold system credentials. Each system is called with the employee's own connection, which expires on a set schedule.	**A missing or expired connection prompts the employee to reconnect. No shared credential stands in.**	Upstream log names the employee and the gateway record names both principals
AA-07	**Response data protection**	Secrets and personal data are removed from tool responses before the agent reads them.	**If filtering cannot complete, the response is blocked by default.**	Rule names and match counts, with matched text omitted
AA-08	**Tool integrity**	Approved tools are monitored for changes to their definitions.	**A changed tool is blocked until an administrator re-approves it.**	Blocked calls recorded against the tool
AA-09	**Prompt protection**	Customer details in prompts are replaced with placeholders, and injected instructions are detected.	**Detected injection is redacted or the prompt is stopped, per policy.**	Prompt record with agent, person and model
AA-10	**Rate limits**	Call limits per person, agent, system or tool, applied together.	**Over budget, the agent receives a tool error with a retry hint.**	Limit events naming the rule, including monitor-only results
AA-11	**Anomaly detection**	Spend, volume, errors and off-hours activity are compared with each person's normal pattern.	**A departure from baseline can deny the next call, or pause, throttle or revoke the instance.**	Anomaly findings per person and agent
AA-12	**Endpoint rules**	A laptop sensor, rolled out with Jamf or Intune, finds agents and local tools and labels shadow AI.	**Unapproved servers and blocked commands are stopped before they run.**	Device events in the Agent Trail under the person
AA-13	**Model access and budgets**	Approved models per team, and spend budgets per team or person with alerts.	**At the budget, requests stop.**	Spend by team, model and person
AA-14	**Decision records**	Each decision is recorded with person, agent, rule and context, and streamed to your SIEM.	**Kept three months in the product, and in your SIEM for as long as your policy requires.**	Records naming person, instance, rule, effect and context

© 2026 SecureAuth Corporation.

CHAPTER 7

Framework alignment

An indicative mapping of the controls in chapter 6 to NIST SP 800-53 Rev. 5 control identifiers and OWASP agentic risk IDs. SecureAuth prepared it to help place agent controls in an existing control library. It is not an attestation of compliance.

ID	CONTROL	SP 800-53	OWASP
AA-01	**Instance binding**	AC-2, IA-2, IA-4	ASI03, ASI10
AA-02	**Client authentication**	IA-5, IA-9, SC-23	ASI03, ASI10
AA-03	**Ordered rules**	AC-3, AC-6	ASI01, ASI03, ASI07
AA-04	**Conditions**	AC-3, AC-6	ASI01, ASI02
AA-05	**Tool classification**	AC-16	ASI01, ASI02
AA-06	**Brokered credentials**	IA-5, SC-28	ASI03
AA-07	**Response data protection**	AC-4	ASI06

ID	CONTROL	SP 800-53	OWASP
AA-08	**Tool integrity**	CM-3, SI-7	ASI04
AA-09	**Prompt protection**	AC-4, SI-10	ASI01, ASI06
AA-10	**Rate limits**	SC-5	ASI02, ASI08
AA-11	**Anomaly detection**	IR-4, SI-4	ASI08, ASI10
AA-12	**Endpoint rules**	CM-7, CM-8, CM-11	ASI02, ASI04, ASI05, ASI10
AA-13	**Model access and budgets**	AC-3, SC-6	ASI08
AA-14	**Decision records**	AU-2, AU-3, AU-6, AU-12	ASI05, ASI07, ASI09, ASI10

Division of responsibility

AREA	AGENT AUTHORITY	ENTERPRISE TEAMS
People and groups	Reads identity and groups from the identity provider at sign-in	Run joiner, mover and leaver processes in the directory
Access decisions	Decides each call against the active rule set	Write and review rules, and approve servers for the catalog
Upstream entitlements	Calls each system as the employee	Maintain permissions in each system of record
Devices	Discovers agents and enforces agent rules	Run device management and host protection
Model use	Redacts prompts, applies model allowlists and budgets	Select providers and evaluate models
Evidence	Records each decision and streams it by webhook	Retain, correlate and alert in the SIEM

References. OWASP Top 10 for Agentic Applications, 2026 · NIST SP 800-53 Rev. 5 · Model Context Protocol specification · OAuth 2.1 and OpenID Connect. A companion whitepaper covers the identity and policy design in technical depth.

Map the controls with SecureAuth engineering
SecureAuth product and engineering leads can map AA-01 to AA-14 to your control library in a working session with your security architecture team.

© 2026 SecureAuth Corporation.