AI SECURITY ENGINEERING FIELD GUIDE · 2026

Chapter 05 · Agent Security

Standalone learning module for LMS delivery and required reading.

FORMAT	USE	SCOPE	AUDIENCE
Standalone PDF	LMS reading module	Single chapter	Learners

CHAPTER 05

Agent Security

Authority

AGENT SECURITY

Map what the agent can read, write, send, execute, purchase, delete, approve, and administer.

> Agent security is delegated action security, not chatbot security.
>
> FIELD GUIDE

> **FIELD GUIDE: AUTHORITY REVIEW**
>
> Map what the agent can read, write, send, execute, purchase, delete, approve, and administer.

This domain covers AI systems that can call tools, browse, query data, update records, send messages, write code, operate workflows, or act across business systems. It matters when model output can trigger state change, external communication, privileged lookup, transaction, workflow execution, or irreversible action.

Review the agent as an actor with delegated authority. The central question is not whether the agent sounds helpful. The question is what it can do when context is hostile, goals are ambiguous, a tool returns misleading data, or approval is rushed.

> **FIELD USE**
>
> Build the authority graph before testing prompts. List each tool, identity, action, target resource, approval gate, log, rollback path, and kill switch.

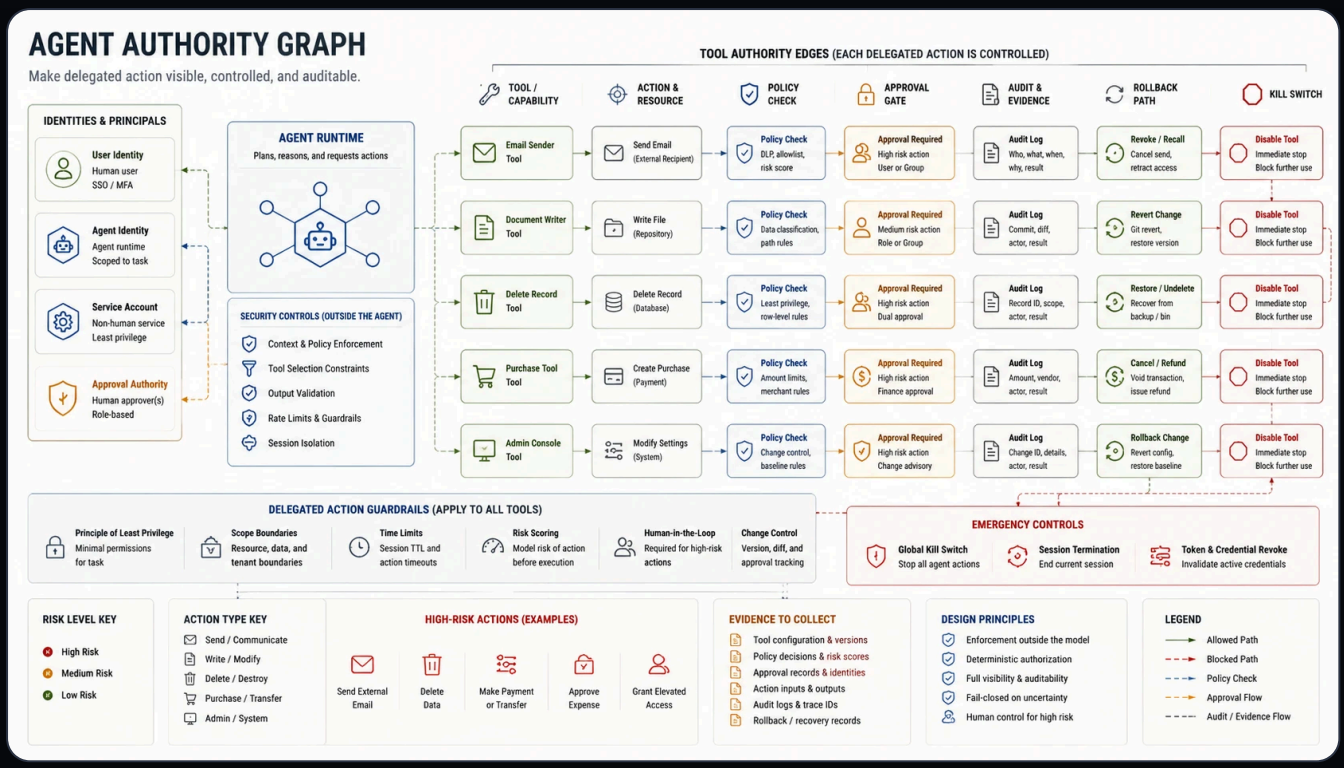

FIGURE 1: FIGURE 6: AGENT AUTHORITY GRAPH MAPPING TOOLS, IDENTITIES, PERMISSIONS, APPROVAL GATES, LOG PATHS, ROLLBACK CONTROLS, AND KILL SWITCHES

WHAT BREAKS HERE?

- Tool permissions exceed the task.
- Approval screens hide the target, authority, data source, or reversibility.
- Tool outputs inject instructions back into the agent.
- Agents chain small actions into high-impact outcomes.
- Logs show final output but not the action path.
- The agent acts through a service account with broader access than the user.

WHAT TO INSPECT

- Tool manifests, scopes, credentials, identities, and resource permissions.
- Approval gates, bypass paths, kill switches, rollback paths, and rate limits.
- Memory, planner, executor, and tool-output handling.
- Cross-system workflows, connector permissions, and admin surfaces.
- Tool-call logs, policy decisions, and action evidence.
- Human handoff, escalation, retry, and scheduled-task behavior.

WHAT TO ASK

- What identity does the agent use for each tool?
- What can the agent do that the user cannot do directly?
- Which actions are destructive, external, financial, privileged, or hard to reverse?
- What context does an approver see before granting action?
- What stops a tool result from becoming a new instruction?
- How can the team pause, revoke, or roll back agent activity?
- What prevents one safe-looking step from enabling an unsafe chain?

WHAT TO TEST

- Excessive permissions against read, write, send, delete, and admin actions.
- Approval bypass, approval fatigue, and incomplete approval context.
- Tool-result prompt injection.
- Action chaining across systems.
- Credential exposure through prompts, logs, tools, or errors.
- Kill switch and rollback execution.
- User-to-agent privilege confusion across tenants, roles, and shared workspaces.

CONTROLS AND GUARDRAILS

- Least-privilege tool scopes by resource, tenant, action, time, and environment.
- Separate user identity, agent identity, service identity, and approval authority.
- Runtime policy engine for tool calls.
- Approval gates with target, action, data source, authority, reason, and reversibility.
- Sandboxing, egress limits, rate limits, kill switches, and rollback procedures.
- Tool-call logging with request, decision, result, approver, and trace ID.
- Step-up authorization for destructive, external, privileged, or irreversible actions.

EVIDENCE TO COLLECT

- Agent authority graph.
- Tool permission matrix.
- Approval design screenshots or records.
- Tool-call policy logs.
- Abuse test results.
- Rollback and kill-switch test evidence.
- Action-chain traces with acting identity and approver identity.

OUTPUT ARTIFACTS

- Agent authority matrix.
- Tool permission review.
- Approval boundary assessment.
- Agent abuse findings.
- Hardening backlog.

RELATED SERVICES AND WORKBENCH TOOLS

TYPE	RELATED PATHS
Workbench	Authority Graph, Adversarial Range, Runtime Proxy
Services	Agentic Workflow Hardening, AI Product Security Assessment, AI Red Team Validation
Handbook	Agentic Permissions, Prompt Injection, Incident Response

DECISION - CONDITIONAL: AGENT LAUNCH DECISION

Do not ship an agent until its authority, approvals, logs, rollback path, and kill switch are visible and testable.

Agent Security AISECURITY.LLC