OWASP GENAI SECURITY PROJECT 1 SEPTEMBER 2026

A new standard for controlling AI agents

- Agent Control Standard v0.1, public preview
- The 2026 Top 10 got the headlines
- This one matters if your agents **act**:
  - move money
  - talk to your customers
  - send data outside your walls

KONSTANTIN TRUNIN

SWIFTWARD.DEV

Two programs and one question

19 HOOKS DEFINED

6 MANDATORY

LOCKED BY THE STANDARD

- the messages: JSON-RPC 2.0 over HTTP(S) or stdio (wire format)
- the five answers (disposition vocabulary)

THE GUARDIAN ALSO EMITS

- Tracing OpenTelemetry, OCSF - optional
- AgBOM CycloneDX, SPDX, SWID - required once a rule checks what the agent is made of

YOURS (DEPLOYMENT-OWNED)

- the engine that runs the rules
- the rules themselves

All 19 hooks, 6 of them mandatory

SESSION

- **sessionStart**
- **sessionEnd**

IN AND OUT

- **userMessage**
  (a person writes to the agent, before the model sees it)

  or **agentTrigger**
  (a schedule, an incoming event, another agent starts the run)
- **agentResponse**
  (the model has produced its reply, before it goes to whoever receives it)

TURN

- turnStart
- turnEnd

TOOLS

- **toolCallRequest**
- **toolCallResult**

KNOWLEDGE AND MEMORY

- knowledgeRetrieval
  (RAG or knowledge base lookup)
- memoryContextRetrieval
  (memory read into context)
- memoryStore
  (memory write)

CONTEXT WINDOW

- preCompact
- postCompact

SUB-AGENTS

- subagentStart
  (a child agent begins)
- subagentStop
  (a child agent ends)

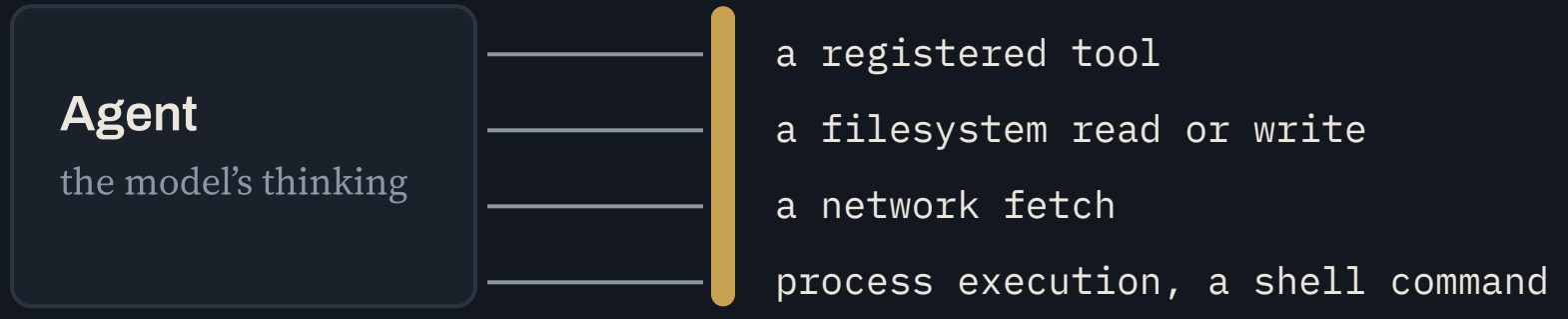
SKILLS

- skillRegister
  (a skill becomes available)
- skillLoad
  (a skill activates in a session)
- skillUnload
  (a skill leaves the active set)

One hook covers everything that leaves the agent

- `toolCallRequest` is **the central enforcement point:** the hook that runs before the action does

EACH OF THESE MUST FIRE `toolCallRequest`



including the operations the framework runs itself

"primitives that bypass `toolCallRequest` are invisible to policy"

The Guardian answers in one of five words

ALLOW — Proceed

the action runs

DENY — Block

a reason is required

MODIFY — Proceed with changes

arguments overridden, fields redacted

ASK — Pause and request approval

the agent waits for an approver on its own side: a person, another agent or a service

DEFER — Verdict not yet reachable

the Guardian holds the case and names how it resolves: more information, a person, or time. On timeout it denies.

Reading the two together: ASK puts the wait on the agent's side, DEFER keeps it inside the Guardian.

Two layers inside the Guardian

DETERMINISTIC LAYER

- always runs first
- a Guardian with only this layer, and no LLM, still meets the standard

AGENT LAYER (THE LLM INSIDE THE GUARDIAN, OPTIONAL)

- runs only when a rule hands the case to it
- gets the request and what the deterministic layer already worked out; the policy code itself stays out of its reach
- its decisions are logged with the model that made them

ON THE AGENT BEING WATCHED:

“The agent MUST NOT have knowledge of hooks”

THE GUARDIAN CHAINS ONE ENTRY PER STEP IT RECORDS

an entry records which step happened; recording what the step contained is optional

What v0.1 leaves open

- Deferred to v0.2:
  - **streaming** - `agentResponse` assumes a finished answer; nothing yet for checking a stream chunk by chunk
  - **A2A wrapping** - MCP is wrapped today (`protocols/MCP/*`); the namespace for agent-to-agent traffic is reserved, its wrapping arrives in v0.2
  - **multi-tenant isolation** - the envelope already carries `tenant_id`, with no isolation rules behind it yet: scoping rules per customer, session and audit come in v0.2
- Conformance is self-declared:
  - no test suite · no registry · nobody to arbitrate a disputed claim

Could your framework fire a hook before every action that leaves your agent?

SOURCES

github.com/GenAI-Security-Project/agent-control-standard
genai.owasp.org/resource/agent-control-standard-acs