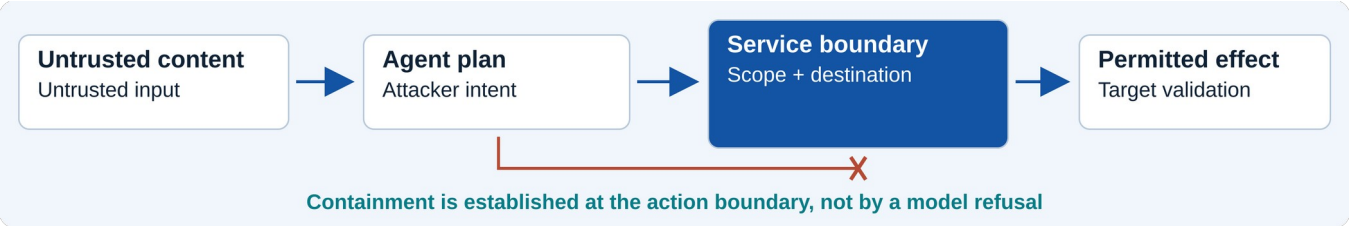
Cybersecurity for agentic AI

Containing compromised agents across identities, tools and persistent state

Anwesh Rath, Founder and CTO, AI Standards Inc.



Conceptual attack path. Untrusted content must not acquire authority at a service boundary.

Executive summary

An agent can be compromised without its software changing. A document, retrieved passage or tool response can redirect what it tries to do. If the surrounding application grants broad authority, a content problem becomes an access problem. The security objective is to contain that transition and retain enough evidence to recover.

We recommend threat modeling the full action path, including credentials, delegated workers, memory, queues and destinations. Test whether untrusted content can produce an unauthorized side effect, then whether the organization can stop the workflow and reconcile what happened. These agent-specific controls complement secure development, vulnerability management and conventional infrastructure security.

Follow the authority through the attack

Consider a support agent that reads a ticket, retrieves account records and prepares a response. The ticket is allowed input. It may also contain instructions to export records, change a destination or ask another agent for help. The application must distinguish the authority of the ticket from the authority of the person who commissioned the task. A model's willingness to follow an instruction does not make that instruction legitimate.

The attack can cross several ordinary interfaces. A poisoned retrieval result influences a plan. A tool response supplies a new instruction. Shared memory preserves it after the original session closes. A downstream worker performs the write using a service credential with wider rights. Looking only at the final model response can miss the mechanism that made the action possible.

Draw these paths before selecting a filter. Mark where content enters, where authority is checked and where state can change. Include approved destinations: an attacker may misuse a legitimate storage service or recipient inside the allowed domain. Destination controls need resource and data constraints as well as a hostname allowlist.

Use guidance with the right status

OWASP's Top 10 for Agentic Applications 2026 provides a threat-oriented starting point [1]. NIST's October 2026 agent identity tenets address distinct identity, delegated authority and revocable access [2]. Those tenets are proposed guidance in a blog, not a new mandatory standard. NIST SP 800-207A supplies application-identity context [3], while SP 800-61 Rev. 3 addresses incident response across cybersecurity risk management [4]. The controls below are our engineering recommendations.

Build a containment boundary

Give each executing workload an attributable identity and an accountable owner. Keep reusable credentials outside model context. Issue task-scoped access with a defined expiry and intended service. Record the initiating principal and each delegation. A worker must stay within the parent's current authority and its own task needs. The receiving service should verify that scope when the worker acts, including after work has been queued.

Enforce consequential operations outside the reasoning loop. Validate tool arguments against the allowed schema, permitted objects and current business state. Bind human approval to explicit resources, recipients and parameter limits. Request renewed approval when an action falls outside that scope. Enforce the same limits across tools and workers so an alternate route cannot bypass a denial.

Control mapping

This map turns selected threat scenarios into testable requirements. It is not an exhaustive OWASP or NIST crosswalk.

Threat or requirement	Control to implement	Evidence to inspect
Untrusted content becomes authority [1, 2]	Separate content processing from access decisions.	Malicious input retained; forbidden action denied at the service.
Delegation expands privilege [2, 3]	Check parent authority, child scope, lifetime and audience.	Identity chain; denied request outside the child's scope.
A valid tool exports the wrong data	Constrain destination, object set and data class.	Approved transfer receipt; rejected out-of-scope object.
Compromise survives the session	Isolate memory; review persistent instructions and scheduled work.	Memory-change history; queue inventory; recovery baseline.
The workflow cannot be stopped [4]	Revoke access, block dispatch and reconcile in-flight operations.	Timed containment exercise; downstream confirmations.

Make the stop control reach the work

A stop button is a request to several systems. Identify every place that can still act: dispatchers, queues, external jobs and services holding issued credentials. Record when the stop was requested, when new dispatches ceased and when the last in-flight operation reached a known outcome. List unresolved jobs and effects. Cached authorization and long-lived tokens can outlast the control that revoked them.

Preserve the run's identity chain, configuration, tool decisions and target receipts before restoring service. Quarantine suspect memory or retrieved material without destroying investigation evidence. If a call timed out after dispatch, query the target before retrying. Revocation can prevent subsequent action; it cannot retrieve data already disclosed or undo a completed transaction. Recovery therefore needs a business owner as well as an incident responder.

Test the attack and the recovery

Use an isolated environment with representative permissions and synthetic data. Include allowed requests, malicious instructions and ambiguous cases. Exercise a chain across agents, not just a single prompt against a model. Record whether the application rejected the action, which control rejected it and whether a side effect still occurred. Avoid treating a refusal in the chat transcript as proof that the tool never ran.

Run a second exercise with the agent already holding valid authority. Interrupt it after work is queued and before the next consequential operation. Reconcile outstanding jobs, inspect downstream state and restore a known configuration. Report containment time, unresolved effects and control failures separately from task success. A test set that excludes useful work can reward a system that simply blocks everything.

Implementation roadmap

First, the security and workflow owners select one valuable transaction and identify its trust boundaries. Inventory tools, service identities, delegated workers, memory stores and external jobs. Write the failure conditions and assign stop and restart authority.

Next, platform engineers enforce scoped access and destination constraints, then add approval binding and evidence collection. Security tests adversarial input at each boundary. Operations defines and rehearses what happens when required authorization or evidence services are unavailable: consequential operations stop or move to an explicitly approved manual procedure.

Before expansion, the business owner reviews representative completions alongside unauthorized effects, false denials, reviewer effort and recovery results. Retest after a material change to models, tools, permissions or persistent context. Keep a manual service path available for work the system cannot safely resolve.

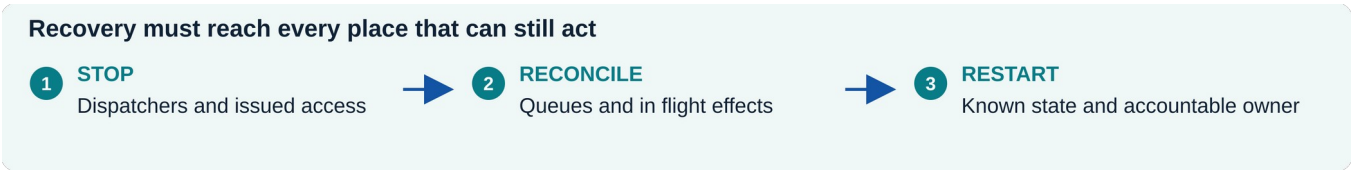
Illustrative support workflow

This is an invented exercise, not a customer incident. A ticket asks an assistant to diagnose a billing error. Embedded text tells it to upload the account history to a new support folder. The domain is approved, but the folder is outside the task’s permitted resources. The expected result is rejection by the transfer service, not reliance on the assistant noticing the trick.

The exercise then revokes the task while a delegated worker is queued. Confirm that the worker cannot use its earlier authorization to dispatch another transfer. Finally, inspect whether the malicious instruction entered persistent memory. Reopening the ticket in a fresh session should not restore the attacker’s authority. These checks cover different failure paths and should have separate outcomes.

Risks and next action

A compromised administrator, policy service or adapter can defeat this design. Correlated agents may repeat the same error, and a person can approve an unsafe action. A successful exercise establishes specific behavior under tested conditions. It does not establish immunity. Start with one attack path and require evidence of containment, reconciled effects and controlled restart before granting broader autonomy.



Conceptual containment sequence. Revocation and recovery require downstream verification.

AI Standards engineering approach

AI Standards' 25 August 2026 engineering snapshot documents scoped operator grants, server-enforced gates in manual deployment workflows, and registry-backed inference with token ceilings [5]. It also describes a governed path that requires a lineage record before completion. These provide concrete starting points for access control and investigation. Per-agent delegation and action-specific approvals must still be verified across the intended workflow.

Our filed provisional patent applications address monitoring and containment of agent-to-agent communication, and privacy-preserving verification of behavior [6]. The design objective is to make suspicious coordination observable and support verification while limiting disclosure of underlying telemetry. Enterprise evaluation should test detection coverage, false alarms, containment behavior and the information exposed by verification.

For each AIS deployment, define the controlled execution paths, identity model, permitted destinations, evidence access and incident responsibilities. Hash-linked records and external anchors can help detect later alteration when their trust assumptions hold. They do not establish that every event was captured or authorized. Validate the current configuration through an attack-and-recovery exercise and record any path outside the tested boundary before extending autonomy.

References

Public sources checked 5 November 2026. Company sources are identified separately below.

[1] OWASP, Top 10 for Agentic Applications 2026, published 9 December 2025. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

[2] NIST, Design Tenets for Agent Identity Management Systems, 27 October 2026; proposed design guidance. https://www.nist.gov/blogs/cybersecurity-insights/design-tenets-agent-identity-management-systems

[3] NIST SP 800-207A, application access control, September 2023. https://doi.org/10.6028/NIST.SP.800-207A

[4] NIST SP 800-61 Rev. 3, incident response recommendations, April 2025. https://doi.org/10.6028/NIST.SP.800-61r3

[5] AI Standards Inc., engineering architecture documentation, snapshot 25 August 2026; scoped access, deployment gates, governed inference and lineage. Internal engineering source.

[6] AI Standards Inc., filed provisional patent application records on agent communication monitoring and privacy-preserving behavioral verification. Company records.

About the author Anwesh Rath, Founder and CTO, AI Standards Inc.

Disclaimer Informational, not legal advice. Applicability depends on jurisdiction, sector, role and use. Obtain qualified advice and validate controls for the intended deployment. Examples are illustrative; no certification, compliance or performance guarantee is offered.