

Preventing Supplier Content from Redirecting Trusted Agent Actions

A Joint Security Test and Response Plan

Phil Hatch | Akholi

October 23, 2026

Outsourcing security managers should test whether untrusted supplier content can cause a trusted agent to act outside its mandate. The test must follow content through retrieval, memory, tool selection, and the final business action.

Test the Boundary Between Content and Authority

A supplier document should supply information to an outsourced agent without acquiring the authority to instruct it. Prompt injection occurs when hostile or misleading content attempts to redirect the agent's behavior. In a two-party service, the buyer may own the affected system while the provider controls how external content enters the agent's context.

The joint security owner should therefore require a demonstrated boundary between business data and authorized instructions. A provider statement that prompts are filtered cannot establish that boundary. Test whether an apparently routine document can change a payment destination, disclose a restricted record, select an unapproved tool, or alter the evidence used to report completion.

The allied May 2026 guidance identifies indirect prompt injection through external sources and the misuse of a trusted agent's privileges. This is an official security recommendation, not proof that a particular supplier is compromised. The proposed response is to test the actual combination of content, privileges, and integrations before consequential use.¹

¹ Australian Signals Directorate's Australian Cyber Security Centre et al., *Careful Adoption of Agentic AI Services* (Australian Signals Directorate, May 1, 2026), "Manage third-party components," "Identity management," "Resilience," and "Human in the loop," <https://www.cyber.gov.au/publication/careful-adoption-of-agentic-ai-services>.

Define a Test That Both Firms Can Interpret

Begin with a named workflow and its authorized outcome. For a hypothetical invoice-matching service, the agent may read invoices and propose matches but cannot change bank details or send records externally. The buyer identifies protected actions and data; the provider identifies retrieval sources, persistent memory, tool descriptions, and downstream agents that could carry the content onward.

Classify test inputs by how they enter the service. A malicious instruction in an invoice differs from one in a tool response, retrieved knowledge article, or earlier memory entry. Record the source, the identity that submitted it, and the point where the system should treat it as untrusted. Testers need written permission and synthetic or appropriately protected data.

Keep the pass condition independent of the agent's explanation. The system passes a specific scenario when the prohibited business action is blocked and a useful record reaches the appropriate owner. An agent that reports refusal while a downstream tool executes the action has failed. A clean narrative is not evidence of containment.

Require Evidence at Each Action Boundary

Figure 1 proposes four test families. Each should include a normal control case to show that useful work still completes. Otherwise, a system that rejects every input could appear secure while failing the service. Agree an acceptable balance through the buyer's risk decision, rather than allow the provider to optimize a security score in isolation.

For each attempt, retain the content identifier, service configuration, input path, granted authority, tool request, enforcement result, downstream state, and alert disposition. Avoid collecting unnecessary sensitive content in the security log. The buyer needs sufficient evidence to reconstruct the attempt without requiring unrestricted access to every provider system.

The study describes visibility limited to each firm's own tools. That mechanism can leave a buyer unable to distinguish a blocked attack from an unobserved action. Its banking accounts motivate a joint evidence test; they do not quantify prompt-injection prevalence or establish attack success in other industries.²

² Phil Hatch, *Agentic AI Outsourcing at Tier 1 Financial Institutions: Full Study Report*, version 1.8.2 (Akholi, forthcoming manuscript dated November 1, 2026), Part 2, sec. 5.1, "Agent Actions Visible Only Inside Each Firm's Own Tools," <https://doi.org/10.2139/ssrn.7164559>.

Figure 1. Test content at the point where it can become action

Test input	Attempted redirection	Required observable result
Supplier file	Change recipient or destination	Action blocked outside mandate
Retrieved content	Override task or policy	Source remains untrusted
Tool response	Call an unapproved service	Tool enforcement denies call
Persistent memory	Repeat instruction after cleanup	Clean run and reconciled queues

Author's proposed operating framework. Adapt to the service and applicable jurisdiction; not an empirically validated standard.

Source: Phil Hatch, proposed framework developed for this paper.

Treat Persistent Context as Part of the Incident

Stopping one run may leave hostile instructions in a memory store, cached retrieval result, or unfinished delegated task. The incident plan should identify every store and work queue that could reintroduce the content. Preserve a protected evidentiary copy before clearing operational context, subject to retention and investigation requirements.

The provider should demonstrate that a fresh run does not recover the poisoned material through another route. The buyer should verify that affected credentials, recipients, and external destinations remain constrained. Clearing a conversational window alone is insufficient when the service retains memory elsewhere or delegates work to another system.

NIST's February 2026 concept-paper announcement includes prompt-injection mitigation among agent identity and authorization questions. It is a project-development initiative rather than a completed certification standard. Managers should use it to sharpen diligence questions while requiring direct evidence from their own service.³

Make Material Changes Reopen the Test

A new retrieval source, model, tool permission, or memory policy can invalidate an earlier result. The joint change record should identify which action boundaries are affected and which

³ National Institute of Standards and Technology, "New Concept Paper on Identity and Authority of Software Agents," National Institute of Standards and Technology, February 5, 2026, project status and identification, authorization, auditing, and non-repudiation questions, <https://www.nist.gov/news-events/news/2026/02/new-concept-paper-identity-and-authority-software-agents>.

scenarios must be rerun. The buyer should retain approval over changes that expand exposure, even if the provider describes them as routine improvements.

Include a recovery test in which the first attempt is blocked but a later routine request retrieves the same hostile instruction from persistent context. The test should establish whether the original content remains reachable and whether the alert identifies the connection. Treat the attempted redirection as test data, not as an instruction to the testers. A recorded failure should identify the affected path and configuration so the provider can correct the mechanism rather than add one phrase to a filter.

Define incident thresholds before the rehearsal. An attempted prohibited action may justify restricted operation and investigation; confirmed unauthorized disclosure or execution needs the applicable incident process. Security, privacy, operational, and contractual notifications may have different recipients and deadlines. The joint incident owner should coordinate them without assuming one notification satisfies all duties.

Mayer Brown's governance analysis supports limiting tool access and defining human checkpoints. The operating method here makes those choices testable between firms: every consequential request must reach an enforcement point that can reject it independently of the agent's interpretation of the document.⁴

Accept the Service on Observed Behavior

The outsourcing security lead should sign a joint test record with the provider before broader deployment. It should identify scenarios executed, failures, limitations, corrective actions, and the configuration tested. Unresolved paths to consequential action should remain disabled or require independent approval until retesting supports release.

Success means the service completes authorized work while hostile content fails to obtain business authority. The continuing obligation is to preserve that boundary as the two firms change the workflow. A recurring test and a workable cleanup process give managers stronger evidence than a general promise that an agent is secure.

⁴ Arsen Kourinian, "Governance of Agentic Artificial Intelligence Systems," Mayer Brown, February 24, 2026, "Mitigation Measures," <https://www.mayerbrown.com/en/insights/publications/2026/02/governance-of-agentic-artificial-intelligence-systems>.

Evidence and Source Note

The study is a nonrepresentative cross-section of 850 vendor delivery-team respondents on 150 engagements serving Tier 1 financial institutions, collected during January–June 2026. Neither banks nor vendors corroborated the accounts. It establishes reported experience, not causation, legal breach, or market prevalence. Its application here is a mechanism to investigate, not an estimate of conditions in other industries. The author-supplied version 1.8.2 is dated November 1, 2026 and is cited as a forthcoming manuscript consulted October 23, 2026.⁵

This is an analytical decision paper based on the supplied study and selected primary public sources checked through October 23, 2026. The original contribution is a joint test record linking hostile content to independently observed business actions and context cleanup. The proposed method has not been independently validated. Legal requirements depend on the entity, activity, jurisdiction, and effective provisions; supervisory observations, research, and professional commentary have distinct authority.

Bibliography

Australian Signals Directorate's Australian Cyber Security Centre et al.. *Careful Adoption of Agentic AI Services*. Australian Signals Directorate, May 1, 2026. <https://www.cyber.gov.au/publication/careful-adoption-of-agentic-ai-services>.

Hatch, Phil. *Agentic AI Outsourcing at Tier 1 Financial Institutions: Full Study Report*. Version 1.8.2. Akholi. Forthcoming manuscript dated November 1, 2026. Author-supplied manuscript consulted October 23, 2026. <https://doi.org/10.2139/ssrn.7164559>.

Kourinian, Arsen. "Governance of Agentic Artificial Intelligence Systems." Mayer Brown. February 24, 2026. <https://www.mayerbrown.com/en/insights/publications/2026/02/governance-of-agentic-artificial-intelligence-systems>.

National Institute of Standards and Technology. "New Concept Paper on Identity and Authority of Software Agents." National Institute of Standards and Technology. February 5, 2026. <https://www.nist.gov/news-events/news/2026/02/new-concept-paper-identity-and-authority-software-agents>.

⁵ Hatch, *Agentic AI Outsourcing*, Introduction, "Executive Summary" and "Evidentiary Basis and Its Limits."