

Test the Agent Configuration You Will Receive

A Critical Infrastructure Acceptance Method for Managers

Phil Hatch | Akholi

October 23, 2026

Infrastructure delivery managers should test the actual supplier configuration in the intended operating context. A benchmark, model label, or demonstration on another setup cannot establish the behavior of the service the operator will receive.

Name the Configuration Under Acceptance

The operator's delivery manager should identify the model, instructions, tools, permissions, data sources, interfaces, and operating conditions proposed for release. These elements collectively determine what the agent can observe and do. A model-version name alone is an incomplete description of the service.

DOE's Stormbreaker work evaluates AI behavior across configurations and conditions rather than certifying a universal capability.¹ The procurement implication is to test the assembled service under the conditions relevant to the operator. Results from another configuration should be treated within their stated scope.

Start with the permitted support task and the consequential actions it must not perform. An agent that prepares maintenance information has a different authority boundary from one that changes a work order or operational setting. The acceptance record should make that boundary visible to engineering, security, and the provider.

¹ US Department of Energy, Office of Cybersecurity, Energy Security, and Emergency Response, *CESER Releases New Testbed to Advance LLM and Agentic AI Evaluation for Critical Infrastructure* (US Department of Energy, Office of Cybersecurity, Energy Security, and Emergency Response, July 16, 2026), configuration transfer and limit discovery, <https://www.energy.gov/ceser/articles/ceser-releases-new-testbed-advance-llm-and-agentic-ai-evaluation-critical>.

Test Both Useful Completion and Forbidden Action

Positive task performance cannot establish that an agent respects its limits. Include attempts to reach prohibited tools, use stale information, act without required authority, and proceed when prerequisites are missing. Observe the receiving system and relevant engineering evidence rather than rely on the agent's report.

The study describes identity and completion-evidence problems in banking delivery.² These are mechanisms to investigate, not infrastructure safety findings. The manager should ask whether its own supplier configuration preserves attributable action and independently verifiable results.

For a hypothetical maintenance-document assistant, successful retrieval should be tested separately from an attempted instruction to alter equipment settings. The service may be useful for the first task while remaining unsuitable for the second. Acceptance should approve the demonstrated scope rather than infer broad operational competence.

Use a Configuration Acceptance Record

Figure 1 proposes four fields: exact configuration, representative conditions, observed behavior, and approved scope. Preserve identifiers and versions sufficient to reproduce or interpret the test. Record differences between the test environment and intended deployment.

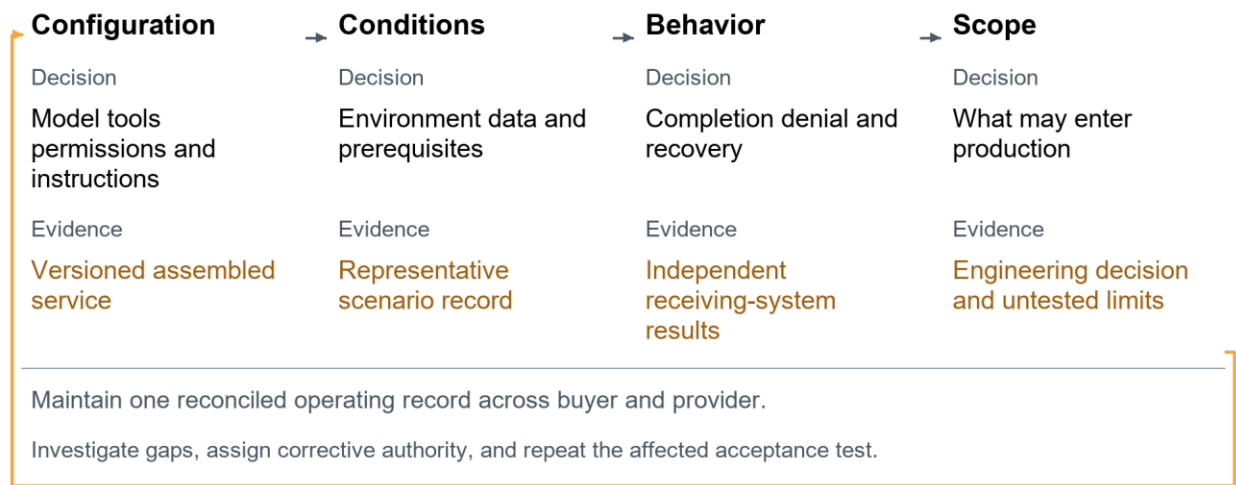
Include missing data, unavailable tools, conflicting sources, delayed responses, and interrupted runs. Engineering should determine which scenarios can be tested safely and which require simulation or another controlled method. The test itself must not create an unauthorized operational exposure.

Allied OT guidance recommends lifecycle governance, testing, and safety-aware failure procedures.³ The manager should connect those disciplines to the local configuration, while recognizing that voluntary security guidance does not replace the operator's engineering or safety obligations.

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 3, Financial Institution Readiness, sec. 4.1, "Ability to Provision Access and Identities for Agents," indicator C108, <https://doi.org/10.2139/ssrn.7164559>; Hatch, *Agentic AI Outsourcing*, Part 3, AI Technology, sec. 4.1, "Accuracy of the Tools' Own Success and Failure Reports," indicator C172.

³ Cybersecurity and Infrastructure Security Agency and Australian Signals Directorate et al., "Principles for the Secure Integration of Artificial Intelligence in Operational Technology," Cybersecurity and Infrastructure Security Agency and Australian Signals Directorate et al., December 4, 2025, secs. 2.4 and 4.2, "Embed Safety and Failsafe Mechanisms," <https://www.cyber.gov.au/publication/principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology>.

Figure 1. The configuration acceptance record



Author's proposed method; adapt to the service and applicable jurisdiction.

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

Keep Human Authority and Recovery Within the Test

Test the conditions under which the agent must abstain or escalate. Confirm the qualified person receives enough evidence and time to decide. A human-review statement is insufficient if the person cannot see the relevant state or restrict the action before it affects the service.

Demonstrate the approved interruption and recovery method. Record outstanding work, accepted actions, and uncertain effects. The operator's engineering function should determine safe withdrawal and resumption. Abrupt network isolation or replaying all tasks may be inappropriate depending on the physical process.

EU, UK, Swiss, US, and other infrastructure requirements vary by sector, site, and activity. The local applicability and safety assessment should determine required evidence and approval. A supplier certificate or successful foreign deployment cannot by itself establish suitability for this environment.

Bind Release to the Demonstrated Scope

A reasonable objection is that exhaustive testing is impossible. Define a proportionate test plan around material actions, uncertainty, and failure consequences. State what remains untested and restrict the scope accordingly. The answer is a bounded acceptance decision, not an unsupported guarantee of complete safety.

Require reassessment when the provider changes the model, tools, permissions, instructions, interfaces, or operating assumptions in a way that affects prior evidence. Preserve the accepted configuration and the route for emergency restriction. A new release should not inherit conclusions that depended on a different setup.

The manager should approve the configuration that has been demonstrated, within the conditions and authority the evidence supports. Useful task performance and enforced limits should both be visible. The service is ready only when the operator can explain what was tested, what was observed, and what work remains outside approval.

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, 2026.⁴

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 configuration-specific infrastructure acceptance record connecting the assembled service, representative conditions, observed completion and denial, and bounded production authority. 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

Cybersecurity and Infrastructure Security Agency and Australian Signals Directorate et al.. “Principles for the Secure Integration of Artificial Intelligence in Operational Technology.” Cybersecurity and Infrastructure Security Agency and Australian Signals Directorate et al.. December 4, 2025.
<https://www.cyber.gov.au/publication/principles-for-the-secure-integration-of-artificial-intelligence-in-operational-technology>.

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>.

US Department of Energy, Office of Cybersecurity, Energy Security, and Emergency Response. *CESER Releases New Testbed to Advance LLM and Agentic AI Evaluation for Critical Infrastructure*. US Department of Energy, Office of Cybersecurity, Energy Security, and Emergency Response, July 16, 2026.
<https://www.energy.gov/ceser/articles/ceser-releases-new-testbed-advance-llm-and-agentic-ai-evaluation-critical>.

⁴ Hatch, *Agentic AI Outsourcing*, Introduction, “Executive Summary” and “Evidentiary Basis and Its Limits.”