

AI Agent Governance Implementation Guide

By Dux Raymond Sy

<http://dux.ai>



The optional companion to the AI Agent Governance Checklist. It explains the checks, shows three practical scenarios and identifies tools that can help. Use the checklist for the team's decision and this guide to investigate or fix a gap.

Observe

Know what is running.

Protect

Secure data, contain harm and recover.

Govern

Assign ownership and enforce permissions, policies and approvals.

Optimize

Prove value and manage cost.

Purpose

Turn an uncertain answer into a specific test or action. Find the relevant control, agree who will handle it and bring the result back to the review.

Who it is for

Process owners, IT and security leads implementing the controls, and business leaders who need to understand what a confirmed answer means. Responsibilities can be handled by your existing team.

When to use it

During setup and testing, when a checklist item is unclear or marked Gap, and after a change or incident. Read the relevant section rather than treating the entire guide as another checklist.

How to use it

1. Find the matching check in Observe, Protect, Govern or Optimize on pages 2-3.
2. Adapt the example to your workflow. Use the scenarios on page 4 or tool references on page 6 when helpful.
3. Reuse existing tests, settings and records. Assign an owner and next action, verify the result, then return to the owner-led checklist review. Partial, failed or unknown evidence for an applicable check means Gap. Use N/A only when the whole check does not apply to the reviewed scope; record why. Complete required approvals.

Tool capabilities depend on the deployment. This guide supports practical review; it is not a complete security assessment or compliance certification.

Apply Observe and Protect

Use the affected check to close a gap before the team review. Business owners set the outcome and acceptable risk. Process owners test the workflow. IT and security verify controls. Reuse existing records; match the effort to the potential harm.

Name the owner before the review

Name one person accountable for the result, plus a support contact and backup. Confirm who can approve use, pause the workflow and authorize restart. Business and data owners approve purpose, permitted use and risk; administrators configure and verify safeguards within their delegated scope. Use existing service ownership or on-call arrangements. Test that a user's report or an operational alert reaches someone who can act.

Owner: Business sponsor.

OBSERVE

01

Check the connections and data

Map each deployed agent to the models, tools or skills, connected data and delegated agents it depends on across platforms. Compare that map with the approved record, including personal accounts. Name data owners; check classification, freshness and oversharing. Fix or exclude stale or misleading sources. Record supplier, version, environment and discovery gaps rather than assuming complete coverage.

Owner: IT or platform owner.

02

Trace an important action or answer

Pick a completed action or important answer. Reconstruct the request, version, sources used, authority or required approval, and result. Protect retained records from the agent's write or delete access. Limit sensitive content, define retention and record logging gaps. Capture observable actions, not hidden reasoning.

Owner: Operations lead.

PROTECT

03

Rehearse stopping and recovery

Stop new, active, queued and delegated work. Test restoring needed data and agent settings to a known-good state; agree acceptable recovery time and data loss. Plan a response for effects that cannot be undone. After a timeout, establish whether the action happened before retrying. Record recovery limits and who authorizes restart.

Owner: Service owner and operations.

04

Test misleading instructions

In an authorized test environment with test data, place hostile instructions in representative documents, web pages or tool responses. Confirm that tested attempts did not expand access or trigger prohibited actions. Record the test and remaining limitations. For code execution, test filesystem and network restrictions outside the prompt.

Owner: Security and engineering.

Apply Govern and Optimize

GOVERN

05

Test access limits

Test an allowed request and one that must be denied, including another user's record where relevant. Bind access to trusted identity, not an identity chosen by the model. Keep delegated work within the same scope. Give permissions an expiry or review date; test revocation when the purpose ends.

Owner: Identity and application owners.

06

Enforce data policies

Check inputs, outputs, memory and logs against approved destinations, provider data-use terms and applicable retention, residency and legal holds. Test enforcement in the application or platform, not just the prompt. Verify deletion or correction where permitted; a legal hold may prevent deletion.

Owner: Data owner with privacy or security.

07

Verify human approval

Separate automatic, approval-required and prohibited actions. Show the reviewer the exact target, recipient, amount and consequence as relevant. Enforce approval where the action executes. Test rejected and expired approvals; a changed action needs fresh approval.

Owner: Process owner and engineering.

OPTIMIZE

Agree the outcome before the review

Agree the job, intended outcome and prohibited actions before testing. Choose a measure such as handling time, accuracy or service quality, and agree the acceptable result before testing. Include human review and error costs.

Owner: Business or process owner.

08

Test quality and business value

Compare realistic results with the current process, including human review, errors and rework. Test missing information, conflicting sources and harmful outcomes. Confirm useful results justify the cost. Uncertainty must reach a person or stop safely; serious consequences require human review.

Owner: Process owner.

09

Test usage limits

Set spending, run-duration and retry limits. Test repeated failure and observe when it stops; record delay or overshoot. An alert is not a spending cap. Assign a budget owner and review cost per useful outcome, not token volume alone. Review adoption, quality and total cost together. Retire unused agents and stale workspaces under the applicable retention rules.

Owner: Platform and budget owners.

Match the review to the risk

SCENARIO 01

Internal knowledge assistant

Answers staff questions from approved internal documents. It cannot change systems or send messages.

Observe. List owned document sources and check freshness and classification.

Protect. Pause unsafe answers; test recovery of approved content and settings.

Govern. Test staff permissions; keep it read-only and route important advice to a person.

Optimize. Measure answer quality and time saved against the current process.

Example decision. A test exposes another team's restricted document. Mark access limits as Gap and hold that connection until access is fixed and re-tested. Draft-only use does not resolve this gap.

SCENARIO 02

Customer service assistant

Drafts replies and proposes refunds from a case record. A person sends the reply and approves the refund.

Observe. Trace each proposed reply and refund to its customer case.

Protect. Contain duplicate requests; rehearse case recovery and response to irreversible errors.

Govern. Test customer separation and approval for the exact recipient and refund amount.

Optimize. Compare resolution time and error costs with the current process.

Example decision. A changed refund amount can still execute without fresh approval. Mark human approval as Gap. Keep refunds disabled until the changed amount requires fresh approval.

SCENARIO 03

IT operations agent

Proposes or executes approved system changes through limited tools.

Observe. Record requested changes, execution results and alert responders.

Protect. Stop queued work and test restoration of known-good configurations.

Govern. Restrict tools and access; test required approval before a production change.

Optimize. Measure successful changes and enforce run and retry limits.

Example decision. Turning off the agent leaves a queued change running. Mark stop and recovery as Gap. Contain that work and re-test the full stop before restoring its authority.

Decide what may proceed

Ask together Any unresolved concern?



Proceed

when applicable checks are confirmed for the stated scope and required approvals are complete. For example, a read-only assistant with verified access, quality and recovery can proceed within that approved boundary.



Limit

use to a smaller scope only after confirming its applicable checks and required approvals. Set a review date. For example, keep a customer agent draft-only with sending and refunds disabled, but only after data access is verified. Draft-only does not fix unauthorized access.



Hold

the affected use while a critical safeguard is unknown or failed, such as unauthorized access, harmful actions, bypassed approval or failed containment and recovery. A smaller scope must not bypass the failed safeguard. Fix, retest and approve restart before resuming held use.

Record the decision

Record the scope, decision owner and review date. Give gaps an owner, action and date; explain N/A. Required organizational, security, privacy or legal approvals still apply. This is not a compliance certification.

Set the next review before proceeding

Set a risk-based review date. Recheck changes to purpose, model, prompt, data, tools, permissions, owner or environment, and after incidents. Monitor abnormal access, actions, quality and cost; route alerts to someone who can act. Keep approval in the existing change record. On retirement, remove access and retain data under policy.

Owner: Business and release owners.

Keep the test record useful

Record the version, date, expected and observed results, and remaining limits. Link existing tests or settings; give gaps an owner and next action.

Choose tools for the gaps you find

Start with tools you already have. These examples support different checks; none replaces ownership, workflow testing or application-level controls. Confirm current licensing, supported environments, configuration and coverage.

AvePoint Trust Platform Observe / Protect / Govern / Optimize Connections, Recovery, Usage, Review	Announced in private preview, with agent discovery, lifecycle governance, security, recovery and cost optimization capabilities. Confirm what is available for your environment before relying on it. Rehearse supported restoration; backup does not mean every agent action can be undone. AvePoint Trust Platform announcement
Microsoft Copilot Studio and API Management Govern / Optimize Connections, Access, Data, Usage	Copilot Studio data policies can group or block connectors. API Management can limit token use for supported model APIs routed through it. Test the deployed path, including delays or temporary overshoot; token limits are not a total spending cap. Microsoft Copilot Studio data policies Microsoft API Management token limits
Google Cloud IAM Govern Access	Grant identities roles on specific Google Cloud resources. Review inherited permissions and test denied access. Authorization inside your application and its external tools still needs separate checking. Google Cloud IAM overview
OpenAI Agents SDK tracing Observe / Optimize Action trace, Quality	Record run events such as tool calls, handoffs and guardrails to investigate behavior. Review sensitive-data collection and retention. Tracing is unavailable under Zero Data Retention; plan evidence collection accordingly. OpenAI Agents SDK tracing
Anthropic Claude Code sandbox Protect / Govern Access, Malicious instructions	Restrict filesystem and network access for sandboxed shell commands. It does not cover every tool: file tools, MCP servers and hooks run outside this boundary. Test the actual configuration and exceptions. Anthropic Claude Code sandboxing

Further guidance

For deeper risk and security review, see [NIST AI RME](#) and [OWASP Excessive Agency](#).

This is a practical review aid, not a complete security assessment or legal compliance certification. Apply your organization's requirements and specialist review where the risk warrants it.