EXECUTIVE BRIEF · NO. 01

Governing agents at work

Business outcomes, accountability and evidence for AI-enabled work.

01

AUTHOR

Przemek Tomczak CPA, CA, CISA

Chief AI Governance Officer, iTmethods

EDITION

September 2026

EXECUTIVE SUMMARY

Executives face pressure to innovate faster and demonstrate a return on AI investment.

When agents can release payments, change records or make commitments to customers, weak controls can expose the organization to financial loss and harm the people it serves. Agents may repeat an error across many transactions before anyone intervenes. The organization remains accountable for the consequences.

Boards and executives need assurance that expected business outcomes are being achieved, risks remain within agreed limits and controls work in practice.

Testing before agents go live is necessary, but it cannot show whether later work keeps meeting expectations. This paper sets out a method for continuous assurance: checking AI-enabled work during operation against intended outcomes and agreed requirements, and using the evidence to decide whether the work should continue, expand, be restricted or stop. It builds on existing management, risk and control responsibilities.

Leaders base those decisions on reports, prepared by management and, where required, by an independent reviewer. Reports should show the work and period covered, what was checked, what failed and what remains unchecked or unresolved. They should identify who assessed the work and the limitations of the conclusion. Management's own checks remain distinct from independent assurance.

The business case and budget should include the people needed to review and correct agent work, with the expertise, authority and time to challenge it and act on problems.

Start with one business process where agents are already in production. Ask the business process owner to answer the questions in **Questions to ask management (C2)** using recent results and supporting evidence. Use the review to decide whether to maintain, expand or restrict agent authority, and what needs to improve.

KEY PANEL

Five questions for reviewing agent work

1. Was the action authorized?
2. Did the safeguards work?
3. Was the intended outcome achieved?
4. What remains unresolved, and what risk remains?
5. What evidence supports the answers?

An authorized action can still produce the wrong outcome, and an apparently good outcome does not establish whether safeguards worked. Use the answers to decide whether the work should continue, expand, be restricted or stop.

SCOPE

Scope and further reading

The method draws on regulation, standards and iTmethods' engineering and client experience. Where it draws on our own work, it describes what that work is designed to do.

This edition reflects regulations, standards and research available in September 2026. Check the current version of any instrument before relying on it.

LIMITS

Using laws and standards

References to laws and standards show where the method can support an obligation or established practice; they do not establish compliance or conformance. Mapping any of it to your own control framework is joint work with your compliance function and counsel, and whether evidence is sufficient for a purpose is a judgment for your own risk, compliance and audit experts.

The main paper covers the decisions and evidence leaders need. For an executive review, start with the executive summary, the five questions and Questions to ask management (C2). Appendix 1 provides control and operating guidance for delivery, risk and audit teams. Appendix 2 covers regulations and standards; the glossary and sources follow.

Contents

A

PART A

Applying controls to agent work

Apply existing controls to the work agents do, and extend them where agents change how that work is done.

RUNNING EXAMPLE

Refunds and fee corrections

At a retail financial-services firm, AI agents handle about 6,000 refund and fee-correction requests a week. They triage each request, decide eligibility, issue refunds of up to $250 on their own, update the customer's record and reply. Larger refunds go to four approvers. The firm and its figures are fictional.

A1 Oversight before and after agents go live

Management defines the intended outcomes, the risk limits that put the board's risk appetite into practice, and accountability for each AI-enabled process. The people responsible need evidence that the work meets those expectations and authority to respond when it does not.

Testing before deployment shows how the system behaved under the conditions tested. Oversight during operation examines the work performed since, including whether the outcome was achieved and whether the controls worked.

WHAT NEEDS ATTENTION	MANAGEMENT RESPONSE
Outputs and chosen actions can vary from case to case	Test representative scenarios, and check whether the business result meets agreed requirements.
Behavior can change when models, instructions, tools, data or suppliers change	Identify material changes and reassess the affected controls and permitted scope.
Agents can take many consequential actions between reviews	Limit cumulative exposure, and define when systems must hold or restrict work.
An agent's self-assessment should not be the only check	Check the work against rules, source records or other evidence.
As agents take the routine cases, reviewers may receive more ambiguous, disputed or consequential cases	Plan and budget the people, skills and time to review, correct and escalate that work.

Figure 1. Agent work and management responses

The plan–do–check–act cycle in Figure 2 connects these responsibilities before and after agents go live.

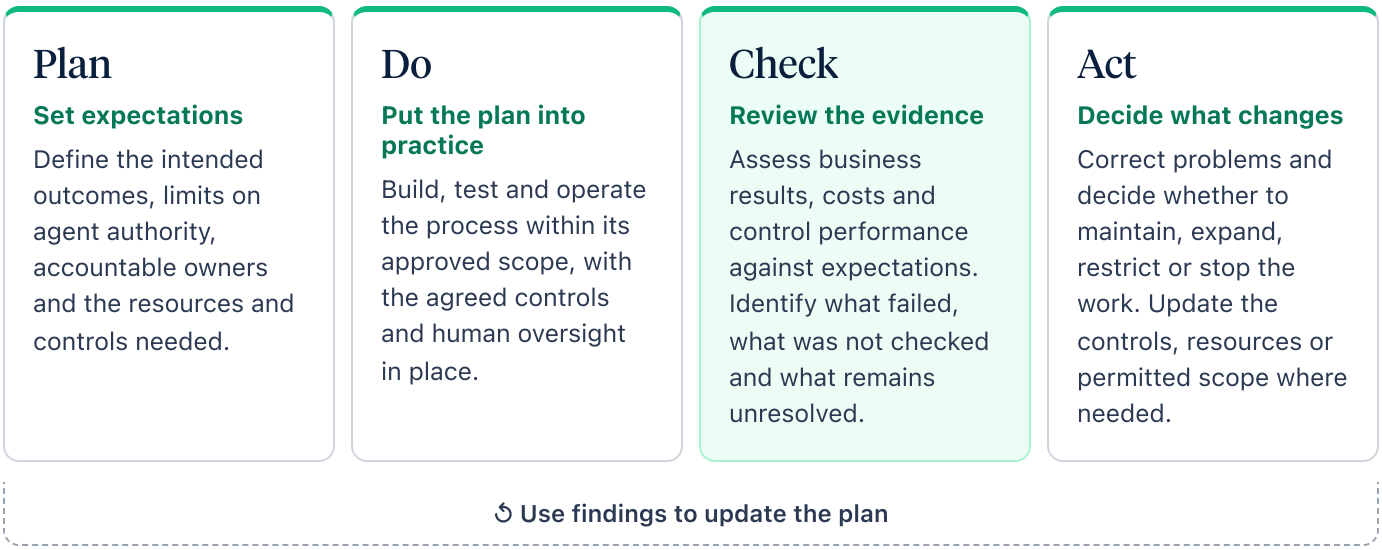


Figure 2. Management oversight of agent-enabled work

How this cycle connects to the detailed workflow: **Appendix 1.1**

Checks run during operation as well as at scheduled reviews. Findings may require immediate intervention or changes to the wider plan. Appendix 1.1 explains how this management cycle connects to the detailed workflow and the frameworks used in this paper.

A2 Control domains for governing agents

Start with the controls the organization already uses. For each AI-enabled process, establish who owns the outcome, what agents may do without approval and which decisions remain with people.

The twelve control domains show where those controls may need extending, from planning and change management to oversight of live work. Use them to identify gaps in responsibility, control and evidence. The detailed reference is in Appendix 1.2.

RUNNING EXAMPLE

The refund process has a business process owner, the head of customer operations, and a record of the three agents involved and the systems each can reach. Its specification defines the outcome: the right amount refunded, the customer's record corrected and an accurate reply.

The $250 limit on refunds an agent may issue alone is one expression of the firm's risk appetite. The process owner also weighs cumulative exposure, customer impact and how quickly errors can be found and corrected.

The twelve control domains

SETS DIRECTION

1 Governance and risk appetite
board and executives · how much an agent may do, at what consequence, before a person decides

↓ sets the limits for every workflow

THE LIFE OF ONE WORKFLOW (the workflow loop of Figure 8)

2	3	4	5	6	7	8
Planning and design business process owner	**Development** engineering	**Verification and validation** risk, engineering	**Deployment and change management** change board, IT	**Operations and monitoring** operations, risk	**Assurance** management checks; risk and compliance oversight; independent internal audit	**Incident and problem management** incident owner, risk

← what operation found, back to the plan · feeds the next plan

Overseeing agents at work (Part B)

FOUNDATIONS EVERY STEP DRAWS ON

9	10	11	12
Training and competence everyone who approves, reviews or owns	**Security, identity and observability** IT, security	**Data governance** data owners, compliance	**Third-party risk management** procurement, vendor risk

Figure 3. The twelve control domains. The detailed reference is in Appendix 1.2.

B

PART B

Overseeing agents at work

During operation, business leaders need evidence that the work remains within the agreed requirements, and a defined response when it does not.

B1 People in the loop are essential

Agents change the work people must do. As routine cases are automated, reviewers may receive a higher proportion of ambiguous, disputed or consequential cases. Plan capacity for that workload, rather than assuming the old mix of cases still applies.

Plan and budget for the people who review, correct and escalate the work as part of the business case. Give reviewers the authority to refuse, the evidence and expertise to judge, and enough time to act. Training, breaks and rotation should reflect the demands of the work.

Set limits on how much work can await review and for how long. Agree what happens when those limits are reached: add capacity, narrow what agents may do, slow the flow or escalate to the business process owner.

Track review delays, overrides, rework and recurring problems alongside speed and cost. To assess whether oversight works in practice, ask reviewers to show a recent case they refused, changed or escalated, and why.

RUNNING EXAMPLE

Four approvers can handle about 300 cases a day, with an agreed queue limit of 600. A billing error increases cases needing approval to 500 a day. As the queue reaches its limit, customers receive a holding reply and the process owner brings in the extra approvers the plan provides for peaks, within the budget in the business case. The process owner then checks whether the added capacity cleared the queue and whether effort and cost stayed within plan.

Appendix 1.3 sets out roles, workload, skills and tools, review limits and measures in detail.

B2 Monitor outcomes and act when things go wrong

Define the intended outcome, operating limits and response to problems before the work begins. Monitor the business results as well as the tasks completed.

When outcomes deteriorate, controls fail or material conditions change, use the agreed response. Systems enforce required holds; authorized people resolve exceptions and decide whether the work should continue, be restricted or stop.

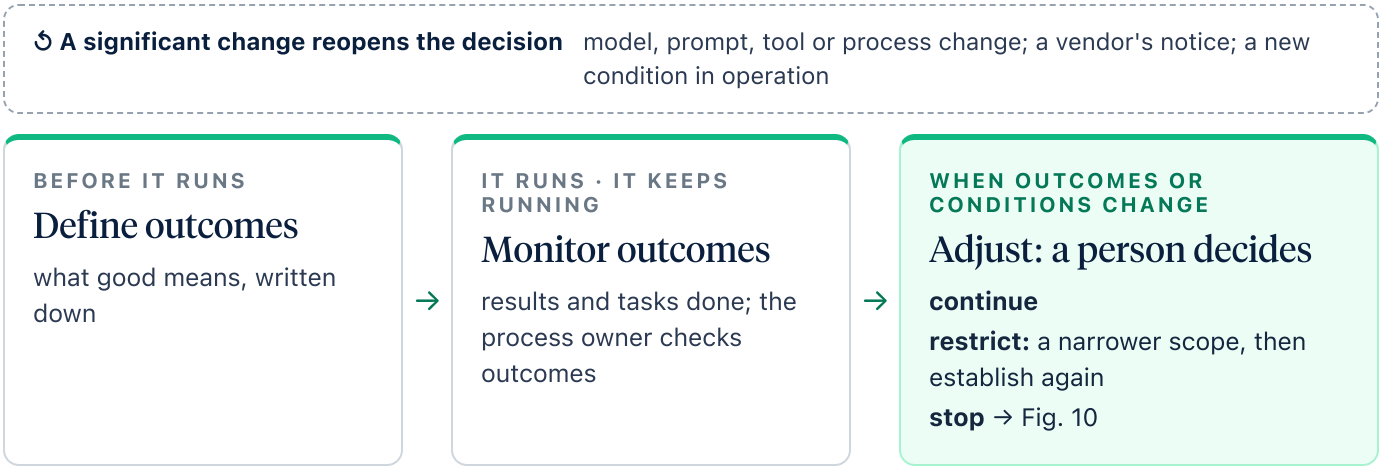


Figure 4. Managing outcomes in operation

Match the response to the potential harm, how quickly it could spread, whether it can be reversed, applicable obligations and the consequences of interrupting the service. A problem in one part of a process may justify restricting that part rather than stopping everything.

If a required pre-action check fails, is inconclusive or returns no result, hold the action for review. If the checking service is unavailable, pause the work or use an approved fallback where the risk permits. Record the evidence gap; an outage is not a passed check.

Problems found after an action require correction and a decision about subsequent work. After a stop, reconcile unfinished work, correct affected records and address the cause before the authorized person approves restart. Preserve a route for affected people to challenge the outcome.

RUNNING EXAMPLE

After a model update, the agent that replies to customers starts promising refunds outside the fee policy. Checks catch the problem. The business process owner restricts the agent's work: staff take over refund-eligibility questions, and the agent keeps handling other replies. Once the fix passes the agreed checks, the process owner returns those questions to the agent.

Appendix 1.5 lists the impact factors and response options, and explains fallback and restart.

B3 Controls in live operation

Enforce limits through the systems agents use. Give each agent identity a named owner and only the access it needs, and test whether the access paths preserve the separation of duties the process requires. A written instruction alone does not prevent an action that the available tools still permit.

Approve and record material changes to models, instructions, tools, data and permissions. In the proposed design, where a consequential action requires approval, the approval identifies that action and the recorded scope it permits. Reassess it when relevant conditions change.

Then check the business result. Authorization does not establish that the work was accurate, complete or achieved its intended outcome. Appendix 1.4 shows how these controls apply to a refund requiring human approval.

B4

THE ROLE OF ASSURANCE

Why it matters for agents

Assurance gives leaders a supported assessment of whether agent work and its controls met agreed criteria within a defined scope and period, so they can decide with confidence how much authority agents should hold. It requires competent, objective reviewers, an agreed assessment process and systems that capture reliable evidence as the work happens.

Three features of agent-enabled work make this important:

1. **They change often.** Models are updated, instructions and tools change and workflows are recomposed, sometimes by a supplier. Testing before an agent goes live shows how it behaved then, not how it behaves now.
2. **They can be wrong in convincing ways.** Agents built on large language models produce likely answers rather than guaranteed ones. They can carry bias, which NIST's AI Risk Management Framework treats as a risk to manage, and in one study, five widely used AI assistants tended to favor a user's stated views over the correct answer.
3. **They cannot be relied on to check themselves.** In one study, language models struggled to correct their own reasoning without outside feedback. Check the work against rules, source records or other evidence instead. A separate checking component helps, but does not by itself make the review independent.

Sycophancy: Sharma et al., *Towards Understanding Sycophancy in Language Models*, arXiv:2310.13548 (2023).

Self-correction: Huang et al., *Large Language Models Cannot Self-Correct Reasoning Yet*, ICLR 2024, arXiv:2310.01798.

Software delivery relies on testing and quality assurance before release. Agents need that discipline carried into live operation: checks on the work itself, as it runs and periodically, matched to the risk (Figure 5). Figure 6 gives examples.

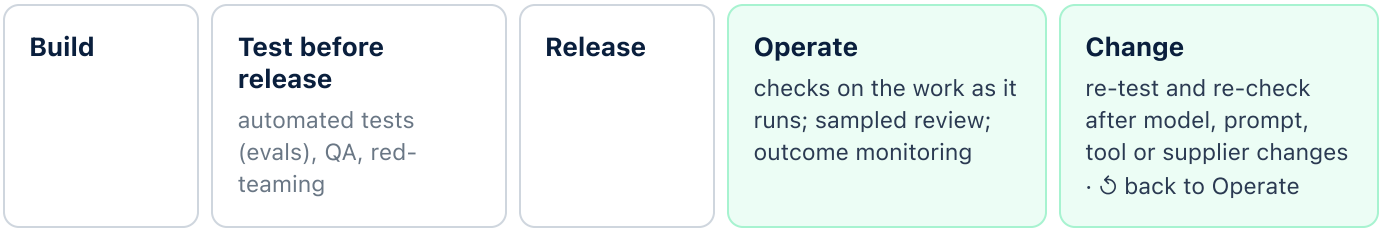


Figure 5. From testing before release to checks in operation. Shaded: the testing discipline carried into live operation.

CHECK	WHAT IT DOES	REFUND EXAMPLE
1 **Input check**	Confirms the request and data are complete and valid before the agent acts	The request names a real transaction on the customer's account
2 **Rule and limit check**	Tests the action against policy and authority limits	The refund is within $250 and the fee policy allows it
3 **Reconciliation**	Compares the result with the system of record	The refund posted matches the approved amount and the right account
4 **Recalculation**	Redoes the calculation separately	The amount is recomputed from the fee schedule
5 **Separate review**	A separate rule set or model reviews the output	A second check reads each reply for promises the policy does not allow
6 **Sampled human review**	A person reviews a sample, weighted to risk	A reviewer examines a weekly sample, weighted to new or unusual cases
7 **Outcome monitoring**	Tracks results over time to spot drift	Reversals, repeat complaints and refund totals by week
8 **Re-check after change**	Re-runs checks after a model, instruction, tool or supplier change	The reply checks re-run on recent cases after a model update

Figure 6. Examples of checks. Appendix 1 sets out the controls for each control domain.

Who provides assurance and how independence is judged: **Appendix 1.7**

Before relying on an assurance conclusion, ask what work and period it covers, what it was measured against, what was checked and what remains unresolved. Management's own checks support that assessment; independent assurance needs a reviewer who is independent of the work. Appendix 1.7 describes assurance more fully, including who provides it and how independence is judged.

B4

What the evidence should show

Keep enough evidence to establish what work was done, what requirements applied, what was checked and what each result meant. Show the population in scope, the coverage achieved and any gaps. Preserve the history needed to understand earlier decisions under the organization's retention requirements.

Keep failed, inconclusive and unassessed work separate: they require different follow-up. Give unresolved issues an owner and a next action.

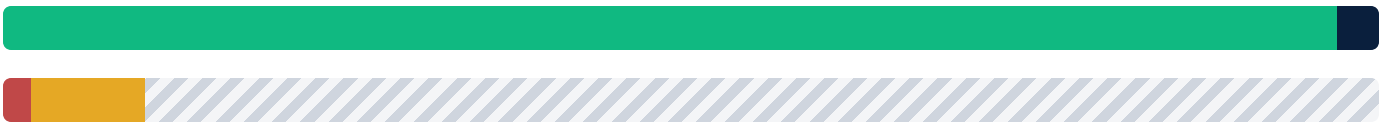
The checks also need scrutiny. Review their test results, false alarms, missed errors, known limitations and failure modes before relying on their findings.

RUNNING EXAMPLE

Last week the process issued 4,800 refunds (Figure 7). A report showing only the 4,655 that passed would hide the rest. The process owner also needs to know the potential impact of unresolved cases and whether corrective actions are overdue.

4,800 refunds issued in the reporting week

the 145 not passed, enlarged below



4,655
passed the checks performed
no exception identified

3
failed
investigate

12
inconclusive
look again

130
not yet checked at the reporting cutoff
reconcile

Figure 7. One week of refund checks. For the records to keep and how to count coverage, see Appendix 1.6.

C

PART C

Oversight and standards

Each organization designs its own controls for AI-enabled work, within the expectations that regulators and standards bodies set. Part C sets out which of them apply to agent work, then gives leaders the questions to ask management about whether the controls exist and work.

C1

Which requirements and guidance apply?

In the frameworks reviewed here, the recurring operating expectations are similar: clear accountability, defined authority and limits, controlled changes, human oversight where required, monitoring of live operation, evidence of what happened, and a response when things go wrong.

First determine, with risk, compliance and legal teams, which requirements apply to the AI-enabled process, given the organization's role and jurisdiction, the system's intended use and classification, and application dates.

They come in four kinds. Binding legal and policy requirements, such as the EU AI Act and, for Canadian federal institutions, the Directive on Automated Decision-Making, apply by role, system and date. Supervisory guidance, such as OSFI's E-23, PRA SS1/23 and the US model-risk guidance, applies to regulated firms within its scope; SR 26-2 expressly leaves generative and agentic AI outside its scope. OSFI's July 2026 technology risk bulletin sets out sound practices institutions can consider. Voluntary frameworks, such as NIST AI RMF and ISO/IEC 42001, name functions and management-system requirements. Professional guidance, such as the IIA's, sets expectations for those who review the work.

The detailed table is in **Appendix 2**. It is a scoping aid, not a conformance assessment: each entry states what the regulation, standard or guidance contributes and what it leaves to you.

C2

Questions to ask management

Use these questions before agents are given more authority and at periodic reviews. Ask management to answer for a specific process, using recent work and supporting evidence.

	EXECUTIVE QUESTION	ASK MANAGEMENT TO SHOW
1	**Is the process delivering the intended outcome at an acceptable total cost?**	Results and actual total cost against the business case, including technology and assurance costs, planned staffing, actual review and correction effort, and material variances.
2	**Who owns the result, and what may agents do without approval?**	The business process owner, the agents and systems involved, enforced limits and decisions reserved for people.
3	**Are access and changes controlled?**	Permissions in use and recent changes to models, instructions, tools or providers, with their impact assessed.
4	**Can people review exceptions competently and in time?**	Capacity, training, queue age and examples of decisions reviewers refused, changed or escalated.
5	**When must we restrict or stop the work, and how do we recover?**	Agreed triggers, decision authority, fallback and recovery arrangements, and routes for affected people to challenge outcomes.
6	**What is wrong, uncertain or unchecked, and what risk remains?**	Coverage for a stated period, separate result categories, unresolved issues and their potential impact, owners and action dates, and any residual risk accepted by the business process owner within their authority.
7	**Who has challenged whether the controls and checks work?**	The reviewer's scope, findings, limitations and independence from the work assessed.

A PRACTICAL NEXT STEP

Choose one process where agents already do consequential work, and ask its business process owner to answer these questions.

Decide whether the intended outcome is being achieved, whether review capacity matches demand and whether the current level of agent authority remains appropriate. Give each material gap an owner, an action and a date.

Use those findings, together with the business case, to decide what should continue, change or stop, and whether wider use is justified.

Checking agent work in live operation, and keeping the evidence that leaders and reviewers need, is the area iTmethods is focused on.

If you are running agents in live operation, we would welcome a conversation.

iTmethods, September 2026.

END OF MAIN PAPER

Appendices 1–4 follow: control reference, regulations and standards, glossary, sources.

APPENDICES · REFERENCE

Appendix 1. Control and operating reference

1.1 Oversight and the workflow · 1.2 Control domains · 1.3 Human review capacity · 1.4 Controls in live operation · 1.5 Responding when problems arise · 1.6 Evidence records and coverage · 1.7 Review arrangements and independence

1.1 Oversight and the workflow

The main paper summarizes management oversight in four steps.

The oversight loop is the organization's management system: plan, do, check and act. It establishes policy and ownership, reviews risk and performance, and directs improvement.

The workflow loop follows one AI-enabled process from planning and design through build, verification and deployment, then operation, monitoring, assurance and response. Evidence from live operation informs decisions about continuing it and about changes to wider governance.

We infer the plan-do-check-act cycle from ISO/IEC 42001's clause structure. The workflow loop adapts the model lifecycle in OSFI's Guideline E-23 and extends it through operation, assurance and response. NIST's Govern, Map, Measure and Manage functions sit across both loops.

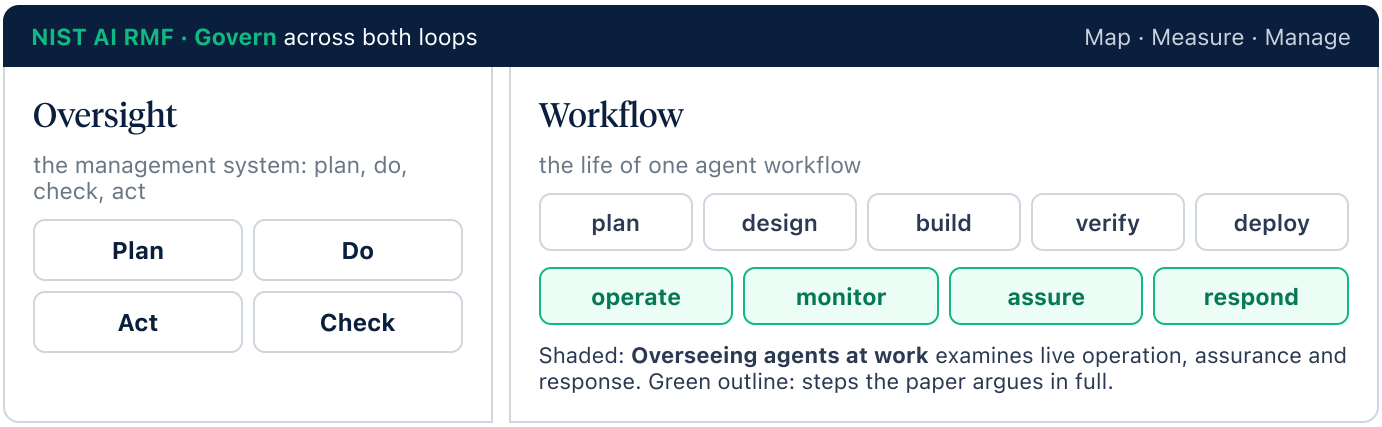


Figure 8. Two loops: oversight and the workflow. Loops read clockwise.

WHAT THE RULES AND STANDARDS PROVIDE

NIST AI RMF is a voluntary framework that sets out four functions for managing AI risk. ISO/IEC 42001 sets out requirements for an AI management system. OSFI Guideline E-23 takes effect 1 May 2027 and covers AI and ML models, including autonomous decision-making. Organizations still need to design the controls for an agent's tools and workflow.

1.2 Control domains

For each control domain, the table names the established control in generally accepted terms, the additional controls we propose for agent work, and the regulations and standards that map to it; Appendix 2 sets out their scope. Use it to check whether the controls you already run cover the work agents now do, and where they need extending.

Items 2 to 6 adapt the AI lifecycle stages in NIST AI RMF 1.0; the other domains group established control disciplines.

#	CONTROL DOMAIN	ESTABLISHED PRACTICE	WHAT WE PROPOSE FOR AGENT WORK	STANDARDS AND REGULATIONS
1	**Governance and risk appetite**	The board sets strategy and risk appetite, and management assigns accountability and keeps an inventory of the AI systems and models in use.	State what the business wants agents to achieve, the consequences it is prepared to accept and who remains accountable for the outcome. Extend the inventory into an agent registry: where each agent is deployed, in which workflows, its owner, and the tools, data and systems it can reach. Translate risk appetite into limits on what agents may do without a person approving each action, and cap the consequential work that may wait for review.	NIST AI RMF GOVERN 1.3 and 1.6; OSFI E-23 (model inventory); PRA SS1/23; HM Treasury Orange Book. OSFI's July 2026 bulletin recommends that institutions consider limits on autonomy.
2	**Planning and design**	Define requirements, acceptance criteria and a business process owner before anything is built.	Document a specification for each agent: intended outcomes, how they will be assessed, and the tools, data and systems it may use. Version the scope each agent is approved to act within, and keep requirements independent of the model under review.	EU AI Act Article 9; OSFI E-23; Canada's Algorithmic Impact Assessment.

Control reference (continues)

#	CONTROL DOMAIN	ESTABLISHED PRACTICE	WHAT WE PROPOSE FOR AGENT WORK	STANDARDS AND REGULATIONS
3	**Development**	Build to the approved design with secure development practices and code review.	Treat prompts, tool definitions, components and settings as code: review, version and pin them. Retain evidence of how each agent was built and of the exact versions released.	EU AI Act Article 17; NIST SP 800-218 and 800-218A; EN 18286 (quality management system).
4	**Verification and validation**	Test against a plan before deployment, with validation independent of development where required.	Evaluate against a documented test plan that records expected results, actual results and the test environment. Include red-teaming and security testing of tools and access. Treat results as evidence about the conditions tested, and keep evaluating after deployment.	NIST AI RMF MEASURE; EU AI Act Article 15; PRA SS1/23. SR 26-2 leaves generative and agentic AI outside its scope.
5	**Deployment and change management**	Approve and record changes, and keep a rollback path.	Track changes to models, prompts, data, tools and permissions separately, because each can change behavior. Approve consequential actions within a versioned scope, one action per approval.	OSFI's July 2026 bulletin recommends that institutions consider enterprise change controls for AI components.
6	**Operations and monitoring**	Monitor live operation against requirements, escalate issues and authorize someone to intervene.	Use current evidence of activity and outcomes to inform a person's decision to continue, restrict or stop the work; for checks required before an action, a failed, inconclusive or missing result holds the action, and findings from completed work require correction and a decision about continued operation. Plan contingencies and mitigations for when live work deviates.	EU AI Act Article 72, the OMB memoranda and Canada's Directive require post-deployment monitoring within their scope and application dates; E-23 and PRA SS1/23 expect it; NIST Manage recommends it; Article 14 adds human oversight.

Control reference (continues)

#	CONTROL DOMAIN	ESTABLISHED PRACTICE	WHAT WE PROPOSE FOR AGENT WORK	STANDARDS AND REGULATIONS
7	**Assurance**	Management's own checks, oversight by risk and compliance, and independent assurance by internal audit (the three lines).	Check live work as it runs and periodically, with checks suited to the risk: for example, validating inputs and outputs against the defined requirements, testing policy and approval rules, reconciling results with the system of record, reperforming a sample of work, and human review where judgment is needed. Retain the evidence so internal audit, compliance and risk can review the checking, rely on it where warranted and evaluate the process itself. Report coverage as counts of what was and was not checked. Keep embedded checks distinct from independent assurance; reviewers need evidence access and stated limits.	IIA Global Internal Audit Standards (Domain I and Principle 7), GTAG 3 and AI Auditing Framework; PRA and US federal guidance support independent review; model-risk validation principles.
8	**Incident and problem management**	Log, investigate and resolve incidents, find root causes, and track corrective action to validated closure.	Distinguish a failed check, doubt, no result and an unavailable checking service. A stop is not a recovery: decide exceptions, reconcile work already started, authorize restart, and give people affected a route to challenge a decision.	EU AI Act Articles 14(4)(e) and 73; Canada's Directive on Automated Decision-Making; GAO Green Book.
9	**Training and competence**	Train people for their roles and assign accountability.	Train and support the people who direct and review agents, and build their expertise in evaluating agents' work. Give them the authority, evidence and time to act on what they find, and treat review capacity as a control limit with a budget.	EU AI Act Article 4 (AI literacy) and Article 14; NIST AI RMF GOVERN 2.2.

Control reference (continues)

#	CONTROL DOMAIN	ESTABLISHED PRACTICE	WHAT WE PROPOSE FOR AGENT WORK	STANDARDS AND REGULATIONS
10	**Security, identity and observability**	Identity and access management, least privilege, segregation of duties, and security logging and monitoring.	Give each agent and tool an owned identity with constrained, short-lived credentials, and keep agents appropriately isolated. Test segregation through the access paths agents actually use, and assess impact by what an agent can reach. Log, monitor and trace agents' actions across the systems they touch.	OSFI B-13; FINOS; OWASP. OSFI's bulletin recommends that institutions consider non-human identities, least privilege and short-lived access.
11	**Data governance**	Govern data quality, lineage, relevance and retention.	Treat inputs, memory and generated outputs as controlled data: inputs can redirect an agent, and memory can be poisoned. Control their origin and lineage, and keep evidence with an owner and a retention rule.	OSFI E-23 Principle 3.2; EU AI Act Articles 10 and 12; OSFI July 2026 bulletin.
12	**Third-party risk management**	Due diligence, contracts, concentration risk, ongoing oversight and audit rights.	Cover model, tool and assurance providers. Reassess when a model or tool provider changes behavior or components. Where a provider does not expose reasoning tokens, that part of its operation is unavailable for capture through the provider's interface.	OSFI's bulletin recommends that institutions consider third-party AI-use notification; NAIC lists contractual audit rights where appropriate and available.

Control reference

1.3 Human review capacity

The work that reaches people changes. A team used to handle a mix of cases, many of them simple. Once agents take the routine cases, what reaches people is mostly what the agents escalate: the unusual, the ambiguous and the disputed, often with a customer already waiting. The remaining cases may take longer and require more judgment, and a full day of them is more tiring than a mixed workload. Plan staffing and support for that changed workload. The risk is that reviewers under that load, and under queue pressure, approve what is in front of them: the tendency the EU AI Act calls automation bias.

Plan the people in the loop like any other capacity:

Roles. Who directs the agents, who reviews their work, who approves consequential actions and who corrects errors. Each needs the authority to say no, enough evidence, the knowledge to judge it and the time to do it.

Load. Size teams for the harder cases that now reach them. Build in breaks, and rotation between demanding and lighter work, so that a reviewer at the end of a shift can still refuse.

Skills and tools. Train people to handle escalated cases and to question an agent's output. Give them tools that help: case summaries that show they were AI-generated, what they drew on and their limits; grouping of similar exceptions; and quick routes to specialists.

Limits. How much work can wait for review, and for how long, before the agents' work is slowed, narrowed or escalated to the business process owner.

Measures. A few that show whether oversight is working: the review queue against its limit, the age of the oldest case, how often people override the agent, and recurring problems.

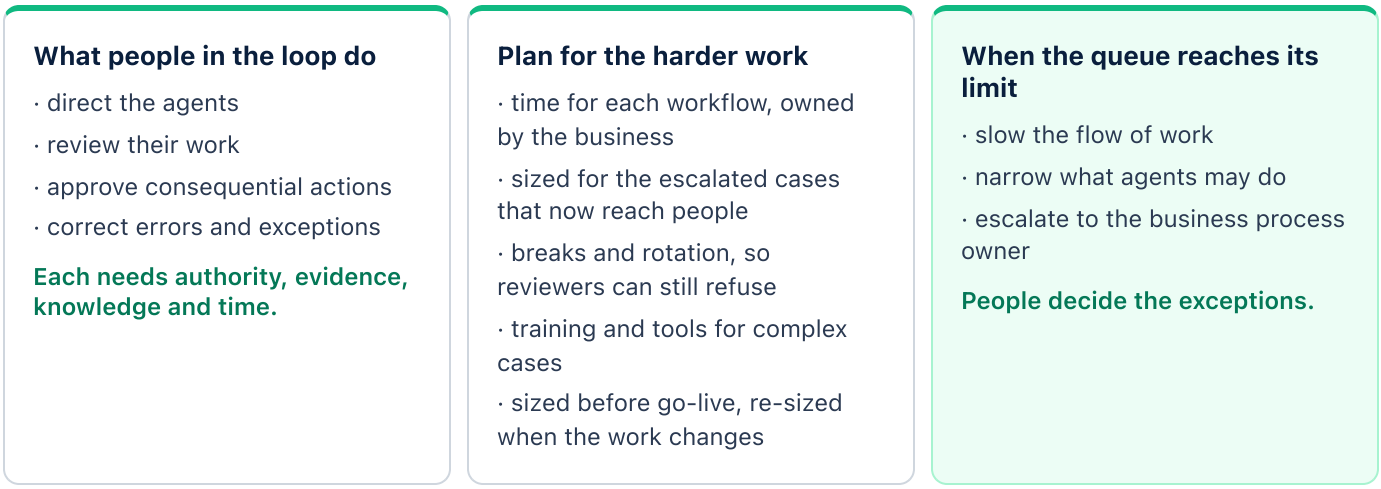


Figure 9. Planning people in the loop

WHAT THE RULES AND STANDARDS PROVIDE

EU AI Act Article 14 sets binding human-oversight requirements, subject to the Act's scope and application dates. Sizing the review capacity for each process is part of the organization's own planning.

1.4 Controls in live operation

The five controls below change most when agents act in live operation. They build on controls most organizations already test, such as accountability, access and change control, and checks on each transaction, rather than adding a separate framework.

1. **Limits on acting alone** (Governance and risk appetite). Define what an agent may do without further approval, including the consequences of its actions, and enforce those limits through the system.
2. **Access and segregation of duties** (Security, identity and observability). Give each agent identity a named owner and only the access it needs. Make sure an agent cannot combine access or actions that would be kept apart for a person, and test this through the credentials and tools the agent actually uses.
3. **Change control** (Deployment and change management). Approve and record changes to models, prompts, tools, data and permissions. Where an action must be prevented, restrict the capability that enables it; a written instruction alone does not enforce that restriction.
4. **Approval for one action** (Deployment and change management). In our design, approval identifies a specific action and a recorded version of its permitted scope. Reassess it when the relevant conditions change; a standing approval may no longer reflect the circumstances of the action.
5. **Checks on completed work** (Assurance). Check whether the completed work was accurate, complete, valid and within the approved scope. Authorization is part of that assessment; it does not establish that the intended business outcome was achieved.

RUNNING EXAMPLE

A customer asks for a $400 refund. The controls apply like this:

1. The refund is over the agent's $250 limit, so the system holds it for a person.
2. The agent can reach only the refunds system, nothing else.
3. Changing the $250 limit takes an approved, recorded change to the refunds system itself.
4. A person signs off this one refund. There is no blanket approval.
5. A later check confirms the right customer received exactly $400, and the record and reply match.

WHAT THE RULES AND STANDARDS PROVIDE

OSFI's July 2026 bulletin sets out sound practices institutions can consider, complementing B-13, B-10 and E-21. FINOS and OWASP provide voluntary control catalogs. Management chooses which of these practices to adopt and how to apply them to its agents' actions.

1.5 Responding when problems arise

Problems come to light in different ways: outcomes fall short, a check fails, a customer complains, a member of staff notices something wrong, or a supplier changes a model the agents rely on. Sometimes the checking service itself stops working, through an outage, a failed change or a supplier problem. When outcomes fall short or something material changes, the person authorized decides whether the work continues, is restricted or stops. The evidence informs that decision; a person makes it.

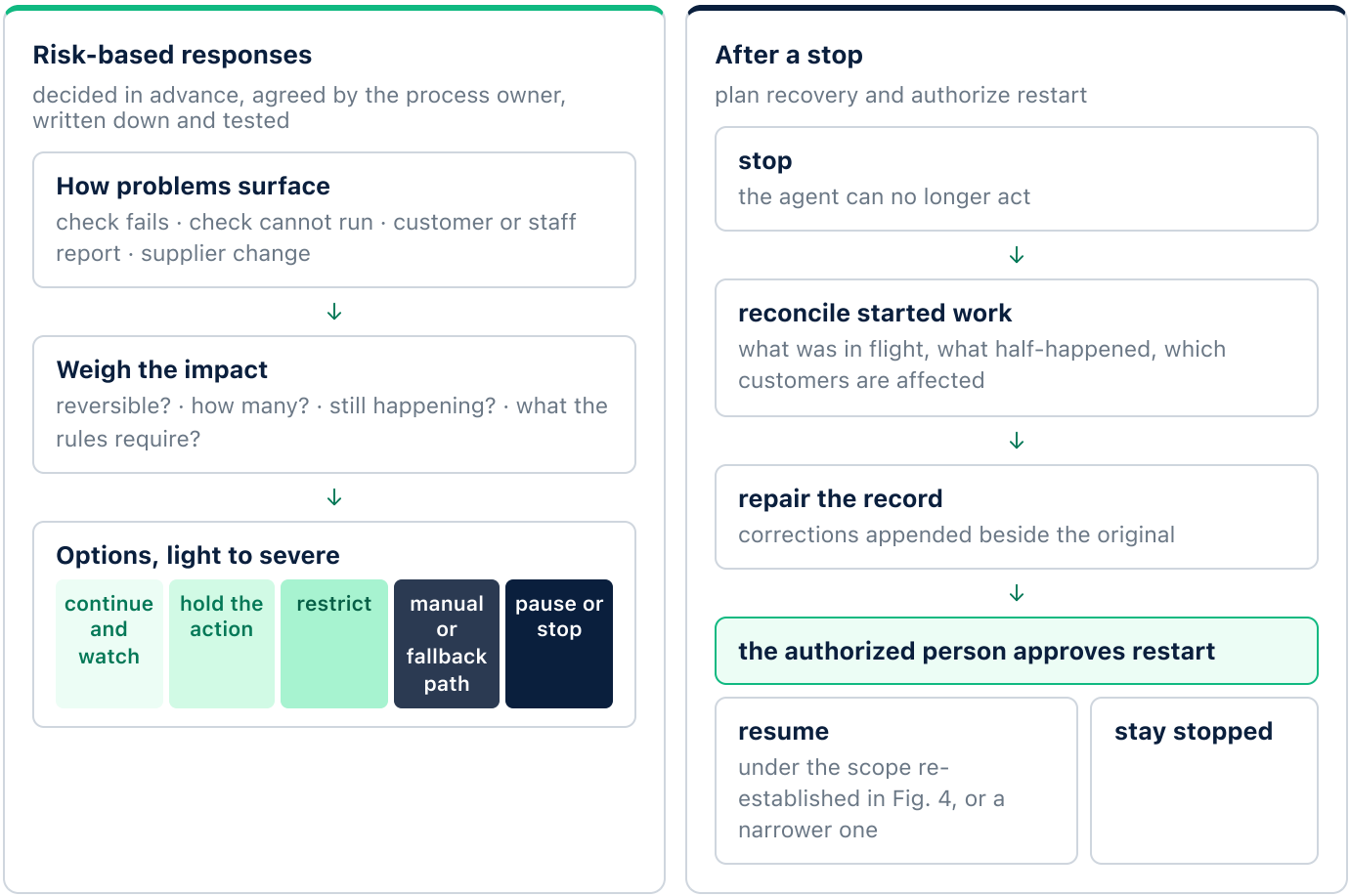


Figure 10. Risk-based responses and restart. Each step has an owner named in the runbook; the customer keeps a route to challenge the outcome.

Base the response on the harm the problem could do to customers and the business. Weigh:

- how much damage one wrong action can do, and whether it can be reversed;
- how many customers or transactions it could reach, and how quickly;
- whether the problem is still happening or was a single event;
- what regulators, customers and contracts require, since some rules specify the response and its timing (see the box below);
- what it costs customers and the business to slow or stop the work.

The options run from light to severe:

Continue and watch more closely. Check more cases, tighten alerts and follow the trend. Suits a contained, low-impact problem.

Hold the action for a person. The system holds the action automatically when a check required before an action finds a problem, cannot reach a conclusion or returns nothing; a person decides how to resolve it.

Restrict. Lower limits, remove tools or send certain cases to people until the cause is fixed.

Use an approved manual or fallback path where the risk permits and the work cannot wait. Record the evidence gap and reconcile the work when checking resumes.

Pause or stop the agents. Use this when harm could spread quickly or cannot be undone.

CONDITION	RESPONSE
Fail; warning, inconclusive, not assessed or abstention; or an empty result	Hold the action for a person, for checks required before an action.
Checking service unavailable	Use the approved degraded or manual path where the risk allows and the work cannot wait, or pause the work; record the evidence gap. The outage does not count as a passed check.

A problem found in completed work calls for correction and a fresh decision on whether the work should continue, since the action has already happened. Distinguish between a check that found a problem and a checking service that was unavailable.

FROM OUR BUILD

In August 2026 a crashed check let a test payment through in our build; we changed it so that a check that crashes holds the action for a person.

In our design, an outage can send an existing process down its approved fallback when the risk allows. That design remains open to review. It should not be read as permission for an agent to continue unchecked.

After a stop, the recovery plan must identify who reconciles work already started, corrects affected records and authorizes restart, accounting for unfinished actions and the people they affect.

An authorized person decides how to resolve each exception; neither the agent nor the checker makes that disposition. The permitted decisions are reject; approve once, with conditions and an expiry; require remediation; narrow scope; escalate; suspend; or another disposition the organization's mandate allows.

Confirm closure through a later check, preferably by someone other than the person who performed the remediation. Where a reviewer concludes that an exception was raised in error, retain the original finding and record it as “disputed as raised”, with the reason.

Record when the failure occurred separately from when it was discovered. Preserve the required route for review or recourse where a decision continues to affect a person.

RUNNING EXAMPLE

One day, the system that checks each refund against the customer's account stops working. The process owner weighs the risk: each refund is small and can be reversed, and customers are waiting. Staff issue 130 refunds by hand, each within the usual limit and marked as not yet checked. When the system is back, the team checks those 130 against the customer accounts, and the process owner restarts normal work.

REFERENCE CASE (CHATBOT)

In *Moffatt v. Air Canada*, 2024 BCCRT 149, the tribunal awarded damages for negligent misrepresentation after a customer reasonably relied on inaccurate chatbot advice. It rejected the airline's denial of responsibility, characterizing that position as treating the chatbot as a separate legal entity. The case illustrates the record a reviewer needs: what the chatbot said, to whom, and when.

WHAT THE RULES AND STANDARDS PROVIDE

EU AI Act Article 14(4)(e) asks for a stop button or a similar procedure allowing a high-risk system to halt in a safe state. Under Article 26(5), a deployer that has reason to consider that a high-risk system presents a risk must inform the provider or distributor and the market surveillance authority without undue delay, and suspend its use; a serious incident must be reported immediately, first to the provider. Article 73 sets deadlines for reporting serious incidents: no later than 15 days after becoming aware, and shorter for the most serious cases. These apply subject to the Act's scope and application dates. Sector rules, such as financial-services incident reporting, may add their own duties. The restart, including reconciling unfinished work, remains a design decision for management.

1.6 Evidence records and coverage

Evidence is what lets someone check, later, whether something is or was true. Figure 11 shows what a reviewer asks about any check, and what to keep to answer each.

QUESTION	WHAT TO KEEP	WHY IT MATTERS
1. What did the check look at?	A record, made when the check ran, of the material it examined	A record rebuilt later shows today's material, which may differ from what the check saw.
2. How much was checked?	Counts of cases in scope, checked and unchecked, for a stated period	A pass rate covers only the cases checked. The unchecked count shows what it leaves out.
3. What did each result mean?	Failed, inconclusive and not assessed, recorded as separate results	Each needs different follow-up: investigate a failure, look again at an inconclusive result, check what was not assessed.
4. What did an earlier decision rely on?	The original record, with any correction dated and added beside it	A reviewer can see what the decision rested on at the time, and what changed since.
5. Can a reviewer check it for themselves?	Enough to follow each check to the same conclusion, and a way to confirm the records are unchanged	The reviewer can then reach their own conclusion from the evidence.

Figure 11. What evidence to keep

The five answers help a reviewer judge whether evidence is relevant, reliable and sufficient, the tests the IIA's Global Internal Audit Standards set for the information internal auditors gather (Standard 14.1).

1. **What did the check look at?** Keep a record made when the check ran, showing the time, the exact material it examined and where that came from, and which version of the check and of the requirement applied.
2. **How much was checked?** For a stated period, keep counts of the cases in scope, how many were checked and which were not. Take the total from the system of record, so it also counts work the check missed. Say when the total is unknown, whether the check covered every case or a sample, and how any sample was chosen. Figure 7 shows one week of refund counts.

3. **What did each result mean?** Keep “failed”, “inconclusive” and “not assessed” apart, as Figure 7 does. Evidence of a breach makes a result a failure; missing evidence makes it inconclusive. For a check before an action, all three results hold the action for review. For a check on completed work, each goes to an owner for correction or further checking.
4. **What did an earlier decision rely on?** Keep the original record and add any correction beside it, with the date, the reason and who made it, so a reviewer can see both. Check that restoring or restarting a system leaves earlier records intact.
5. **Can a reviewer check it for themselves?** Keep enough for a competent reviewer to follow what each check did and reach the same conclusion. A check that uses a model may answer differently when run again, so keep the answer it gave. Give the reviewer the records as an evidence package, with an integrity check they can run independently of the team that produced it, to confirm the records are unchanged. Integrity checks do not establish that the underlying work was correct or the evidence complete.

Name the owner who applies the organization's retention and deletion requirements to these records, rather than leaving either to the tool's default. Keep the material a record identifies for as long as the record itself, within data protection limits, so a reviewer can still follow the check.

WHAT THE RULES AND STANDARDS PROVIDE

EU AI Act Article 12 sets a binding technical logging requirement for high-risk systems, subject to the Act's application dates. Article 26(6), on the same dates, requires deployers to keep the logs those systems generate automatically, to the extent the logs are under their control, for a period appropriate to the intended purpose of the high-risk AI system, of at least six months, unless provided otherwise in applicable Union or national law, in particular Union law on the protection of personal data. An evidence package supports your obligations under logging and record-keeping duties, and is neither an attestation nor a certification. Beyond these logs, the organization determines what evidence its checks keep and who owns it.

1.7

Review arrangements and independence

> The IIA's Global Internal Audit Standards (2024) glossary defines assurance: "Statement intended to increase the level of stakeholders' confidence about an organization's governance, risk management, and control processes ... when compared to established criteria." It defines assurance services as "Services through which internal auditors perform objective assessments to provide assurance."

What assurance involves

Assurance means two things. It is a conclusion: a statement of how far leaders can rely on what was done, judged against suitable criteria. It is also the work that produces that conclusion: examining and testing what actually happened against those criteria.

Delivering assurance takes four things together:

People. Competent, objective reviewers: management checks the work, risk and compliance oversee it, and internal audit or an external provider reviews it independently.

Process. Criteria suited to the risks and objectives, a defined scope and period, and an agreed method for testing the work and recording the conclusion.

Technology. Systems that capture evidence as the work happens, keep it intact and let a reviewer reach it.

Evidence. Relevant, reliable and sufficient records, each showing where it came from and what it covers.

Management's own checks support an assurance assessment. Independent assurance needs a reviewer who is independent of the work and reaches their own conclusion from the evidence. Management, risk and compliance, and the independent reviewer each test the evidence before relying on it.

Embedded controls and independent assurance are distinct disciplines. An embedded checker can be useful and still not be independent. An operational checking service used by management remains part of management's controls, even when an external supplier operates it. An external assurance engagement is a separate arrangement, assessed against its mandate and independence requirements. Internal audit can provide independent assurance where it is positioned independently, with the chief audit executive reporting directly to the board, as can an external assurance provider within its mandate and independence requirements.

The reviewer weighs all the evidence for relevance, reliability and sufficiency: the operator's own records, model-provider tests and platform controls, each with its origin and scope visible. The reviewer also needs evidence about the checker itself: test results, false alarms, known failure modes and the classes of problem it has not seen. In our own testing, all findings in one overnight run traced to one checker bug.

Internal audit can rely on management's checks and on other assurance providers once it has evaluated them: who did the work, how independent and competent they are, how they did it, and whether the evidence supports their results. It records the basis for that reliance. The reviewer must assess the evidence and reach their own conclusion about what it supports. Each finding then needs an owner and a date for correction.

RUNNING EXAMPLE

As part of its risk-based audit plan, internal audit reviews the refund process and its controls. For a sample of refunds, it checks that each was properly authorized, correctly calculated, backed by evidence and paid to the right customer. It then reports, for a stated period, whether the controls were designed well and worked as intended. The weekly report is one input; the conclusion is internal audit's own.

WHAT THE RULES AND STANDARDS PROVIDE

PRA SS1/23 and OSFI E-23 expect validation independent of model development within their scope. US federal guidance also requires uninvolved review in specified cases. SR 26-2 expressly leaves generative and agentic AI outside its scope. PRA SS1/23 has no equivalent exclusion: at firms within its scope, its model definition and validation principles must be considered when assessing an agent's underlying models. The organization decides where independent review of agent work is required and who can perform it.

APPENDIX 2

Regulations and standards in detail

Requirements and guidance: check scope and limits.

The Section column shows where Appendix 1 discusses each one: Oversight and the workflow (1.1), Control domains (1.2), Human review capacity (1.3), Controls in live operation (1.4), Responding when problems arise (1.5), Evidence records and coverage (1.6), and Review arrangements and independence (1.7).

REGULATION OR STANDARD	WHAT IT CONTRIBUTES IN OPERATION	SCOPE AND WHAT IT LEAVES TO YOU	SECTION
NIST AI RMF 1.0 voluntary, 2023	Govern, Map, Measure, Manage; inventory; monitoring; deactivation.	Voluntary. Defines functions rather than a specific operating-evidence design.	1.1 1.2
ISO/IEC 42001:2023 certifiable management system	Management-system clauses covering policy, lifecycle, data and third parties.	A system standard: decisions on individual actions stay with you. Described here at clause-title level. iTmethods holds no ISO/IEC 42001 certification.	1.1
COSO internal control and ERM 2013 / 2017	Control environment, risk assessment, control activities, information and monitoring.	Principles-based: management designs the controls. Described here through COSO-based US federal standards.	1.2
EU AI Act Regulation (EU) 2024/1689, as amended by Regulation (EU) 2026/1744	Risk management, data, logging, human oversight, robustness, quality management, deployer duties, post-market monitoring and incident reporting. Penalty provisions applied from 2 August 2025 and transparency duties from 2 August 2026; Annex III high-risk obligations apply from 2 December 2027; product-embedded high-risk from 2 August 2028.	No separate agent category. Application depends on role, system classification, intended use and the Act's phased dates.	1.3 1.5 1.6
OSFI Guideline E-23 effective 1 May 2027	Model inventory, risk-based intensity, traceable data, validation, monitoring and residual model risk.	Uses model-risk vocabulary. Controls for an agent's tools and workflow require a separate assessment.	1.1 1.2

(continues)

REGULATION OR STANDARD	WHAT IT CONTRIBUTES IN OPERATION	SCOPE AND WHAT IT LEAVES TO YOU	SECTION
OSFI technology risk bulletin July 2026 sound practices	Lifecycle ownership, autonomy limits, non-human identities, change control, prompt/output monitoring and third-party AI-use notification.	Sound practices institutions can consider; complements B-13, B-10 and E-21 and points to E-23 for model risk.	1.2 1.4
US SR 26-2 and 2026 interagency model-risk guidance supervisory	Effective challenge by objective experts and a model inventory.	SR 26-2 explicitly excludes generative and agentic AI from its scope by footnote, applying instead to traditional statistical and quantitative models and non-generative, non-agentic AI. It directs your own risk management to cover what it leaves out.	1.7
UK PRA SS1/23 supervisory	Model identification, governance, development and use, independent validation and mitigants.	Scope depends on the firm and models used. The text does not use AI-specific terminology, while its “dynamic models” definition reaches systems that adapt autonomously.	1.7
Canada Directive on Automated Decision-Making binding for federal institutions	Impact assessment, scaled requirements, human decision at higher impact levels, recourse and training.	Applies to federal government automated decision systems within its scope.	1.2
US OMB M-25-21 and M-25-22 binding for US federal agencies	Chief AI officer and governance board, use-case inventory, signed risk acceptance, independent review and appeal.	Federal agencies only; confirm the memoranda's current status before relying on them.	1.2
FINOS, OWASP, CSA and IIA voluntary / professional	Least privilege, per-action authorization, decision records, security practices and internal-audit guidance.	Useful catalogs and professional guidance, but not binding requirements.	1.2 1.4 1.7

APPENDIX 3

Glossary

Agent	Software that can choose steps or tools and take business actions.
AI-enabled work	Business work in which an AI system decides, recommends or acts.
Assurance	Both a conclusion and the work that produces it: a statement, measured against agreed criteria, of how far leaders can rely on a piece of work, and the examining and testing behind it (after the IIA Global Internal Audit Standards). Independent assurance needs reviewers independent of the work and of the checks built into it.
Continuous assurance, as used in this paper	Checking AI-enabled work against its intended business outcome and agreed requirements, using the evidence to inform decisions on whether the work continues and within what limits.
Coverage	The work assessed over a stated period, with the in-scope population and selection method made clear. State when the total population is unknown.
Failed / inconclusive / not assessed	Failed: examined, and the evidence shows the requirement was not met. Inconclusive: examined but unresolved, for example because evidence was missing. Not assessed: not examined.
Exception and disposition	An issue raised by a check, and the decision an authorized person makes about it.
Evidence package	A bundle of evidence and verification material prepared for review.
Reperformance	Running a check again. It is distinct from an independent re-execution by an outside party.
Non-human identity	The credential or workload identity an agent uses, with a named owner.

APPENDIX 4

Sources

REGULATION AND SUPERVISORY GUIDANCE

European Parliament and Council of the European Union. *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. OJ L, 2024/1689, 12 July 2024. eur-lex.europa.eu

European Parliament and Council of the European Union. *Regulation (EU) 2026/1744 amending Regulations (EU) 2024/1689, (EU) 2018/1139 and (EU) 2023/1230 (Digital Omnibus on AI)*. OJ L, 2026/1744, 24 July 2026; entered into force 27 July 2026. eur-lex.europa.eu

Office of the Superintendent of Financial Institutions (Canada). *Guideline E-23 — Model Risk Management (2027)*. Published 11 September 2025; effective 1 May 2027. osfi-bsif.gc.ca

Office of the Superintendent of Financial Institutions (Canada). *Guideline B-13 — Technology and Cyber Risk Management*. Effective 1 January 2024.

Office of the Superintendent of Financial Institutions (Canada). *Generative and Agentic Artificial Intelligence: Implications for Technology, Cyber Security, and Operational Resilience*. Technology Risk Bulletin, July 2026. osfi-bsif.gc.ca

Board of Governors of the Federal Reserve System. *SR 26-2: Revised Guidance on Model Risk Management*. 17 April 2026. federalreserve.gov

Board of Governors of the Federal Reserve System, Federal Deposit Insurance Corporation and Office of the Comptroller of the Currency. *Supervisory Guidance on Model Risk Management*. SR 26-2 attachment, 17 April 2026.

Prudential Regulation Authority. *SS1/23 — Model risk management principles for banks*. Supervisory statement, current version 16 April 2026 (updating 17 May 2023). bankofengland.co.uk

Treasury Board of Canada Secretariat. *Directive on Automated Decision-Making*. Published 29 March 2021; page modified 24 June 2025. tbs-sct.canada.ca

US Office of Management and Budget. *M-25-21 — Accelerating Federal Use of AI through Innovation, Governance, and Public Trust*. 3 April 2025. whitehouse.gov

US Office of Management and Budget. *M-25-22 — Driving Efficient Acquisition of Artificial Intelligence in Government*. 3 April 2025. whitehouse.gov

National Association of Insurance Commissioners. *NAIC Model Bulletin: Use of Artificial Intelligence Systems by Insurers*. Adopted 4 December 2023. content.naic.org

STANDARDS AND FRAMEWORKS

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1, January 2023. nist.gov

National Institute of Standards and Technology. *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities.* NIST SP 800-218, February 2022.

National Institute of Standards and Technology. *Secure Software Development Practices for Generative AI and Dual-Use Foundation Models: An SSDF Community Profile.* NIST SP 800-218A, July 2024.

The Institute of Internal Auditors. *Global Internal Audit Standards.* Issued 9 January 2024; effective 9 January 2025. Glossary; Standard 14.1. theiia.org

The Institute of Internal Auditors. *The IIA's Artificial Intelligence Auditing Framework.* 2024.

The Institute of Internal Auditors. *Global Technology Audit Guide (GTAG) 3: Coordinating Continuous Auditing and Monitoring to Provide Continuous Assurance.* 2nd edition, March 2015.

The Institute of Internal Auditors. *The IIA's Three Lines Model: An Update of the Three Lines of Defense.* Position paper, 2020; updated September 2024. theiia.org

US Government Accountability Office. *Standards for Internal Control in the Federal Government (Green Book)*. GAO-25-107721, May 2025.

HM Treasury. *The Orange Book – Management of Risk – Principles and Concepts.* 2023 edition.

International Organization for Standardization and International Electrotechnical Commission. *Information technology — Artificial intelligence — Management system.* ISO/IEC 42001:2023, first edition, December 2023. iso.org

Committee of Sponsoring Organizations of the Treadway Commission (COSO). *Internal Control — Integrated Framework.* May 2013. coso.org

Committee of Sponsoring Organizations of the Treadway Commission (COSO). *Enterprise Risk Management — Integrating with Strategy and Performance.* June 2017. coso.org

CEN-CENELEC Joint Technical Committee 21. *Artificial intelligence — Quality management system for EU AI Act regulatory purposes.* EN 18286:2026, available 22 July 2026. cencenelec.eu

OWASP Gen AI Security Project. *OWASP Top 10 for Agentic Applications for 2026.* Published 9 December 2025.

FINOS (Fintech Open Source Foundation). *FINOS AI Governance Framework.* Version 2, 20 October 2025.

Cloud Security Alliance, AI Safety Initiative. *The AI Agent Governance Gap: What CISOs Need Now.* 3 April 2026.

RESEARCH ON LANGUAGE MODELS

Sharma et al. *Towards Understanding Sycophancy in Language Models.* arXiv:2310.13548, 2023. arxiv.org

Huang et al. *Large Language Models Cannot Self-Correct Reasoning Yet.* ICLR 2024; arXiv:2310.01798. arxiv.org

CASES

Civil Resolution Tribunal of British Columbia. *Moffatt v. Air Canada.* 2024 BCCRT 149, February 2024. civilresolutionbc.ca

EXECUTIVE BRIEF · NO. 01

Governing agents at work

Business outcomes, accountability and evidence for AI-enabled work.

ABOUT THE AUTHOR

Przemek Tomczak, CPA, CA, and CISA

Chief AI Governance Officer, iTmethods. Technology executive and former chief auditor.

ONLINE EDITION

itmethods.com/resources/executive-briefs/governing-agents-at-work

This edition reflects regulations, standards and research available in September 2026. Check the current version of any instrument before relying on it. References to laws and standards do not establish compliance or conformance.

iTmethods | Reign

Enterprise AI. Governed.