AI GOVERNANCE RESOURCE

AI Agent Audit Program & Risk Checklist

A structured audit program mapping AI/agent failure modes to test steps, anchored to ISO/IEC 42001, NIST AI RMF 1.0, and the EU AI Act.

This document is part of AI Agent Audit Program series. Use alongside Companion Guide for CAEs.

Prepared by: Ishan Jain | ishan@jainishan.com | jainishan.com

Version 1.0 | June 2026

Table of Contents

1. Purpose, Scope & How to Use This Program

1.1 Purpose

This audit program gives Internal Audit a repeatable, standards-anchored method for assessing risk in AI agents and AI-enabled decision systems. It is designed to move AI assurance work beyond informal opinion or vendor marketing claims, and onto a documented control framework that a CAE can defend to the Audit Committee, regulators, and external assessors.

It is built as a control matrix: each row links a known AI/agent failure mode to (a) the governing clause(s) in a recognized standard or regulation, (b) specific audit test steps, and (c) the evidence an auditor should request to substantiate the control.

Use it (*alongside AI Agent Audit Program - Companion Guide for CAEs*) to plan engagements, build work programs in your GRC/audit tool, and document conclusions on a per-control basis.

1.2 Standards Anchoring

Three frameworks are used throughout, referenced by short code:

- **ISO/IEC 42001:2023** (AIMS) — the AI Management System standard. Clause and Annex A references follow the published structure (Clauses 4–10; Annex A controls A.2–A.10).
- **NIST AI RMF 1.0** — referenced by function: GOVERN, MAP, MEASURE, MANAGE, including the Generative AI Profile (NIST AI 600-1) where relevant to agentic/GenAI behavior.
- **EU AI Act (Regulation 2024/1689)** — referenced by article where a binding legal obligation exists (primarily for high-risk systems under Title III, GPAI obligations under Title V, and transparency duties under Art. 50).

Where an organization is not directly subject to the EU AI Act, treat those references as leading practice benchmarks — many EU AI Act control expectations (risk classification, human oversight, logging, post-market monitoring) are converging with what regulators elsewhere (UK, US sectoral regulators, Canada, Singapore) are beginning to expect.

1.3 Scope of “AI Agent”

For this program, an “AI agent” is any system that uses a model (LLM or otherwise) to perceive context, make a decision, and take an action — including calling tools/APIs, executing code, retrieving or writing data, or initiating transactions — with limited or no per-action human approval. This includes:

- Autonomous or semi-autonomous LLM agents (tool-calling, multi-step planning, RAG-based retrieval agents)
- Embedded AI features inside SaaS/ERP/CRM platforms that take automated action (e.g., auto-categorization, auto-approval, auto-remediation)
- Co-pilot/assistant tools where output materially influences a downstream decision without independent review
- Third-party or vendor-hosted agents acting on company data or systems

1.4 How to Use This Program

1. Scope the engagement: identify which agents/use cases are in scope using the inventory test in Section 3 before testing controls — you cannot audit what is not inventoried.
2. Risk-rank the population: prioritize agents with transaction authority, access to sensitive data, or customer/regulatory exposure (see Risk Tiering, Section 2.3).
3. Walk each domain: each domain is a self-contained control matrix. Select rows relevant to the agent(s) in scope; not every control applies to every agent.
4. Document evidence and exceptions in the working paper template (Section 11).
5. Roll up findings using the rating scale in Section 12 for committee reporting.

Three Lines of Defense Assumption

This program assumes a three-lines-of-defense model: business/product owns the agent (1st line), AI risk/compliance sets policy and reviews (2nd line), Internal Audit independently tests (3rd line). Where the 2nd line function does not yet exist, flag this as a governance gap — its absence is itself a finding.

2. Risk Taxonomy & Materiality

2.1 Failure Mode Taxonomy Used in This Program

Seven domains organize the failure modes:

- **Governance & Accountability** — ownership, policy, lifecycle management
- **Excessive Agency & Authorization** — scope of autonomous action, privilege, blast radius
- **Prompt Injection & Adversarial Exposure** — manipulation of agent behavior via untrusted input
- **Human Oversight & Escalation** — meaningful human control, override, and stop mechanisms
- **Data Lineage, Quality & Privacy** — training/grounding data provenance, retention, rights
- **Vendor, Model & Third-Party Risk** — supply chain, model change management, concentration risk
- **Monitoring, Logging & Incident Response** — observability, traceability, post-deployment monitoring

2.2 Materiality Drivers

When scoping depth of testing, weight materiality by:

- Degree of autonomy (read-only vs. write/transact vs. irreversible action)
- Data sensitivity (PII, financial, health, confidential/IP, regulated data)
- Decision impact on individuals (EU AI Act Annex III triggers: employment, credit, insurance pricing, law enforcement, education access, essential services)
- Reversibility and blast radius if the agent acts incorrectly at machine speed and scale
- Regulatory exposure (sector-specific rules, cross-border data transfer, consumer protection)

2.3 Risk Tiering

Tier	Profile	Audit Implication
Tier 1 – Critical	Autonomous transaction authority, irreversible actions, EU AI Act high-risk classification, or material financial/customer impact.	Full-scope annual test of all seven domains; independent technical testing (e.g., red-team/injection testing) required.
Tier 2 – Elevated	Write access to internal systems, recommendation/decision-support with human-in-the-loop, moderate data sensitivity.	Targeted testing on Sections 5–8 every 12–18 months; design-only review of Sections 4, 9, 10.
Tier 3 – Limited	Read-only assistant/co-pilot, no system write access, low-sensitivity data, human reviews all outputs before use.	Self-assessment by 1st/2nd line with periodic IA sample-testing (e.g., 1 in 3 audit cycles).

3. Pre-Engagement: AI Agent Inventory & Population Test

Standards basis: ISO 42001 Clause 4.1/4.2 (context), Clause 6.1 (risk identification), Annex A.3 (AI system inventory); NIST AI RMF MAP 1–2; EU AI Act Art. 6 & Annex III (classification duty).

Before testing any control, Internal Audit must independently verify the completeness of the AI agent inventory — management's stated population is itself an audit object, not a given.

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
No central inventory of AI agents / shadow AI	*ISO 42001 Annex A.3; NIST MAP 1.1; EU AI Act Art. 6*	**1.** Request the organization's AI/agent system inventory and risk classification register. **2.** Independently corroborate via procurement records (SaaS contracts referencing AI/agent features), cloud billing (LLM API usage, vector DB spend), and IT asset/SaaS discovery tools. **3.** Interview engineering, data science, and business unit leads to identify agents built or embedded outside the formal inventory ('shadow AI'). **4.** Reconcile discrepancies; sample at least one unlisted agent (if found) and trace it through governance gates retroactively.	• Inventory/register export • Procurement & vendor contract list (AI clause search) • Cloud/API billing detail (token usage by service) • SaaS discovery tool output	**H**
Risk classification not performed or outdated	*EU AI Act Annex III; ISO 42001 6.1.2; NIST MAP 1.5*	**1.** For each in-scope agent, confirm a documented risk classification exists (e.g., minimal/limited/high-risk equivalent, or EU AI Act Annex III mapping). **2.** Test whether classification was re-performed after material changes (new tool access, new model, new use case) — request change log. **3.** Assess classification criteria for consistency across business units (look for inconsistent self-rating to avoid high-risk obligations).	• Risk classification worksheets • Model/agent change log • Classification methodology document	**H**

4. Governance & Accountability

Standards basis: ISO 42001 Clauses 5 (Leadership), 6 (Planning), 7 (Support), 9 (Performance Evaluation); NIST AI RMF GOVERN 1–6; EU AI Act Art. 9 (risk management system), Art. 17 (QMS for providers).

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
No defined AI system owner / accountable executive	*ISO 42001 5.3; NIST GOVERN 2.1*	**1.** Confirm a named, accountable owner exists for each Tier 1/2 agent (not just a team). **2.** Verify the owner has authority to pause/decommission the agent without multi-week escalation. **3.** Check RACI or governance charter for AI explicitly assigns 2nd-line review responsibility distinct from the building team.	• AI governance charter • RACI matrix • Org chart with AI accountability mapped	**M**
AI policy does not address agentic/autonomous use cases	*ISO 42001 5.2, Annex A.2; NIST GOVERN 1.1*	**1.** Review the AI acceptable use / governance policy for explicit coverage of autonomous action, tool-use, and agent-to-agent interaction (not just 'chatbot' use cases). **2.** Test whether the policy was updated within the last 12 months or after the last major agent capability change. **3.** Confirm policy has been formally approved by a body with authority (risk committee, board sub-committee) and communicated to relevant teams.	• AI policy document with version history • Approval minutes • Training/communication records	**M**
No pre-deployment risk assessment / AI impact assessment gate	*ISO 42001 6.1.2, 8.2; EU AI Act Art. 9, Art. 27 (FRIA); NIST MAP 1–5*	**1.** Select a sample of deployed agents and request the pre-deployment risk/impact assessment for each. **2.** Verify assessment occurred BEFORE production deployment, not retrospectively (check dates against deployment logs). **3.** For Tier 1 agents, confirm a Fundamental Rights Impact Assessment equivalent was performed if the use case touches Annex III categories. **4.** Test that the assessment includes failure mode analysis specific to agentic behavior (not a generic model card).	• Pre-deployment assessment documents • Deployment timestamps/change tickets • FRIA or equivalent, if applicable	**H**
Lifecycle/change management gaps (model swaps, prompt changes, tool additions not re-assessed)	*ISO 42001 8.3, Annex A.6; NIST MANAGE 2.3*	**1.** Obtain the change log for a sample of agents over the audit period. **2.** Identify changes to underlying model, system prompt, tool/plugin access, or permissions scope. **3.** Test whether each material change triggered a re-assessment or only a code review. **4.** Confirm rollback procedures exist and were tested at least once in the period.	• Change/version log • Re-assessment records tied to specific changes • Rollback test evidence	**M**

5. Excessive Agency & Authorization

Standards basis: ISO 42001 Annex A.6 (Operation), A.8 (System Design); NIST AI RMF MANAGE 1–2, GenAI Profile (NIST AI 600-1) — action/agency-specific controls; EU AI Act Art. 14 (human oversight), Art. 9 (risk management).

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
Agent operates with standing/broad credentials rather than least-privilege, task-scoped access	*ISO 42001 Annex A.6; NIST MANAGE 1.3; EU AI Act Art. 14(4)*	**1.** Obtain the IAM/service-account configuration for the agent (API keys, OAuth scopes, DB roles). **2.** Compare granted permissions against the documented minimum permissions required for the agent's stated task. **3.** Test for over-provisioning: can the agent read/write tables, call APIs, or access systems unrelated to its function? **4.** Confirm credentials are scoped, rotated, and not shared with human user accounts or other agents.	• IAM policy/role export • Service account permission list • Task specification document for comparison	**H**
No tiered autonomy / approval thresholds for high-impact actions	*NIST MANAGE 2.2; ISO 42001 Annex A.6; EU AI Act Art. 14*	**1.** Identify which actions the agent can take fully autonomously vs. which require human approval (e.g., transaction value thresholds, irreversible deletes, external communications). **2.** Test a sample of high-impact action types: confirm an approval gate actually fires in a non-production test, not just in design documentation. **3.** Verify thresholds are risk-based (e.g., dollar amount, customer impact) and were approved by risk/compliance, not solely by engineering.	• Action authorization matrix • Approval gate test/demo evidence • Threshold approval sign-off	**H**
Tool/function-calling surface is broader than necessary (agent can call destructive or sensitive tools 'just in case')	*NIST MANAGE 1.1, 1.4; ISO 42001 Annex A.8*	**1.** Request the full list of tools/functions/APIs exposed to the agent's orchestration layer. **2.** Map each tool to a business justification; flag any tool (e.g., raw SQL execution, file system delete, payment initiation) not directly tied to the agent's core task. **3.** For any agent with code-execution capability, confirm sandboxing/isolation controls (containerization, network egress restriction) are in place and tested.	• Tool/function registry • Sandbox/isolation architecture diagram • Penetration or sandbox-escape test results	**H**
Agent-to-agent or agent-to-agent-orchestrator chains compound privilege without re-evaluation (privilege escalation across a multi-agent pipeline)	*NIST MANAGE 1.2; ISO 42001 Annex A.8*	**1.** For multi-agent/orchestrator architectures, map the full call chain (orchestrator → sub-agents → tools). **2.** Test whether a downstream sub-agent inherits the orchestrator's full privilege set or is independently scoped. **3.** Trace one end-to-end transaction through logs to confirm the documented architecture matches actual runtime behavior.	• Architecture/data-flow diagram • End-to-end trace logs for a sampled transaction • Sub-agent permission scoping documentation	**M**
Kill switch / immediate suspension capability untested or absent	*EU AI Act Art. 14(4)(e); ISO 42001 Annex A.6; NIST MANAGE 2.4*	**1.** Confirm a documented mechanism exists to immediately disable the agent's ability to act (not just delete the chat session). **2.** Request evidence the kill switch was tested in the audit period (tabletop or live test). **3.** Verify time-to-disable against an internal SLA; assess whether the SLA is adequate given the agent's blast radius.	• Kill-switch runbook • Test/exercise log with timestamps • SLA documentation	**H**

6. Prompt Injection & Adversarial Exposure

Standards basis: NIST AI 600-1 (GenAI Profile) — direct/indirect prompt injection controls; ISO 42001 Annex A.8 (robustness), A.10 (incident management linkage); EU AI Act Art. 15 (accuracy, robustness, cybersecurity).

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
No defense against indirect prompt injection via retrieved content (documents, emails, web pages, tool outputs)	*NIST AI 600-1 §2.9; EU AI Act Art. 15*	**1.** Confirm whether the agent ingests untrusted content (web search results, customer emails, third-party documents, scraped pages) as part of its context. **2.** Request evidence of input sanitization, content-source tagging (trusted vs. untrusted), or instruction-hierarchy enforcement between system prompt and retrieved content. **3.** Where available, request results of adversarial/red-team testing specifically targeting indirect injection; if none exists, this is a finding for any Tier 1/2 agent. **4.** If feasible within audit authority and a controlled non-production environment, observe or commission a basic injection test (e.g., a planted instruction in a test document attempting to override agent behavior).	• System architecture showing data sources fed to the agent • Red-team/injection test report • Input handling/sanitization design doc	**H**
No segregation between system instructions and user/external content in the model context	*NIST AI 600-1 §2.9; ISO 42001 Annex A.8*	**1.** Review prompt construction/orchestration code or design documentation to determine whether system instructions are structurally separated from user and tool-retrieved content (e.g., via distinct message roles, delimiters, or a guarded instruction layer) rather than concatenated as plain text. **2.** Test whether the agent's exposed tools can be invoked purely from content embedded in a document or webpage without a corresponding legitimate user request.	• Prompt/context construction design doc • Code walkthrough notes (no need to extract full prompt text)	**H**
Output is trusted/executed without validation (agent acts on its own output, e.g., generated code or commands, without a checkpoint)	*NIST MANAGE 1.4; ISO 42001 Annex A.6*	**1.** Identify any agent workflow where model output is directly executed (code execution, command construction, API call parameters) rather than treated as a recommendation. **2.** Test for output validation: schema checking, allow-listing of permitted actions/parameters, or static analysis prior to execution. **3.** Confirm logging captures both the generated output and the validation decision (pass/block) for traceability.	• Output validation logic/design doc • Sample of blocked vs. executed actions from logs	**H**
No process to track and respond to emerging prompt-injection / jailbreak techniques	*NIST GOVERN 4.1; ISO 42001 9.1, 10.1*	**1.** Ask who is responsible for monitoring emerging adversarial techniques (internal threat intel, vendor advisories, OWASP LLM Top 10 tracking). **2.** Request evidence of at least one instance where a new technique was assessed against the organization's agents in the audit period. **3.** Confirm a feedback loop exists from security/AI red-team findings into agent guardrail updates.	• Threat intelligence process documentation • Guardrail update change log tied to a specific threat	**M**

7. Human Oversight & Escalation

Standards basis: EU AI Act Art. 14 (human oversight, binding for high-risk systems); ISO 42001 Annex A.6, A.9; NIST GOVERN 3, MANAGE 2.2–2.4.

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
Human oversight is nominal/illusory ('rubber stamp' approval) rather than meaningful	*EU AI Act Art. 14(1)-(3); ISO 42001 Annex A.6*	**1.** Sample a population of human-reviewed agent decisions and assess average review time per decision relative to decision complexity. **2.** Interview reviewers to assess whether they have sufficient information, context, and time to genuinely evaluate (vs. a default-approve UI pattern). **3.** Test approval-rate statistics: a near-100% approval rate with minimal review time is a red flag for rubber-stamping. **4.** Confirm reviewers are not measured/incentivized purely on throughput in a way that discourages pushback.	• Review time logs • Approval/rejection rate statistics • Reviewer interview notes • Reviewer performance metric documentation	**H**
Reviewers lack competence/training to oversee AI-generated outputs	*EU AI Act Art. 14(3)(b); ISO 42001 7.2 (Competence)*	**1.** Request training records for staff designated as human-in-the-loop reviewers for AI agent outputs. **2.** Confirm training covers known failure modes (hallucination, bias, injection indicators) specific to the agent's domain, not generic AI awareness training. **3.** Test recency: training completed/refreshed within the policy-defined interval.	• Training curriculum and completion records • Competency assessment results, if any	**M**
No clear escalation path when the agent encounters low-confidence, novel, or out-of-policy scenarios	*NIST MANAGE 2.4; ISO 42001 Annex A.6*	**1.** Confirm the agent has a defined behavior for low-confidence or out-of-scope situations (escalate to human, decline to act, flag for review) rather than defaulting to a best-effort guess. **2.** Sample logged instances of escalation/refusal events and assess whether they were appropriately handled and within SLA. **3.** Test edge cases (where audit access allows) to confirm escalation logic fires as designed.	• Escalation policy/logic documentation • Sample of escalated cases with resolution • Edge-case test results	**M**
Automation bias not addressed — humans defer to agent output without independent judgment	*EU AI Act Art. 14(2); NIST GOVERN 3.2*	**1.** Interview a sample of human overseers about how often and why they have overridden the agent in the past quarter. **2.** A near-zero override rate across many decisions over time is itself evidence worth probing — corroborate with case-file review of a sample of approvals. **3.** Confirm UI/UX design does not pre-bias the human toward acceptance (e.g., one-click approve vs. equally accessible reject/modify options).	• Override rate statistics • UI/workflow screenshots • Sample case files with override rationale	**M**

8. Data Lineage, Quality & Privacy

Standards basis: ISO 42001 Annex A.7 (Data for AI systems); NIST MAP 2, MEASURE 2.2; EU AI Act Art. 10 (data governance), GDPR alignment where applicable.

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
No documented data lineage for training, fine-tuning, or RAG-grounding data	*ISO 42001 Annex A.7; EU AI Act Art. 10(2)-(3); NIST MAP 2.3*	**1.** Request data lineage documentation for any data used to train, fine-tune, or ground (RAG) the agent's outputs. **2.** Trace a sample data source back to its origin: confirm rights to use (license, consent, contractual basis) are documented. **3.** Test for inclusion of any prohibited or out-of-scope data categories (e.g., special category personal data without lawful basis).	• Data lineage/catalog documentation • Data rights/licensing records for sampled sources • Data classification tags	**H**
Sensitive/regulated data flows into agent context without adequate controls (PII in prompts, logs, or vector stores)	*GDPR Art. 5/32 (where applicable); ISO 42001 Annex A.7; NIST MAP 2.2*	**1.** Map data flows: what categories of data are passed into the agent's context window, retrieved via RAG, or written to vector/embedding stores. **2.** Test whether PII/sensitive data is masked, tokenized, or filtered before reaching the model, where required by policy. **3.** Inspect a sample of logs and vector store entries for unredacted sensitive data at rest. **4.** Confirm data retention periods for prompts/logs/embeddings align with policy and any applicable regulatory retention limits.	• Data flow diagram • Masking/redaction configuration • Sample log and vector-store extract (reviewed under appropriate access controls) • Retention policy and deletion logs	**H**
No process to assess or monitor training/grounding data quality, bias, or drift	*ISO 42001 Annex A.7; NIST MEASURE 2.6, 2.11*	**1.** Confirm a documented process exists for assessing data quality and representativeness before use, and periodically thereafter. **2.** For agents informing decisions about individuals, request evidence of bias/fairness testing across relevant demographic dimensions, where lawful and applicable. **3.** Test whether data drift monitoring exists for RAG knowledge bases (stale/contradictory source documents) and whether alerts have been acted upon.	• Data quality assessment reports • Bias/fairness testing results, if applicable • Drift monitoring dashboard/alert log	**M**
Cross-border data transfer via model/vendor infrastructure not assessed	*GDPR Ch. V (where applicable); ISO 42001 Annex A.7; EU AI Act Art. 10*	**1.** Identify where model inference and any data storage (logs, embeddings) physically/jurisdictionally occur for each vendor in scope. **2.** Confirm appropriate transfer mechanisms (SCCs, adequacy decisions, regional hosting commitments) are documented for any cross-border flow. **3.** Cross-check against the vendor contract's data residency clause to confirm actual practice matches contractual commitment.	• Vendor data residency documentation • Transfer mechanism/legal basis records • Vendor contract data-residency clause	**M**

9. Vendor, Model & Third-Party Risk

Standards basis: ISO 42001 Annex A.10 (Third-party and supplier relationships); NIST GOVERN 6 (third-party); EU AI Act Art. 25 (provider/deployer obligations split), Title V (GPAI provider obligations).

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
No due diligence performed on model/agent vendor prior to deployment	*ISO 42001 Annex A.10; NIST GOVERN 6.1*	**1.** Request the vendor due diligence assessment (security, AI-specific risk, sub-processor disclosure) performed before onboarding. **2.** Confirm the assessment addresses AI-specific risk (model provenance, training data practices, known incident history) and is not a generic vendor security questionnaire. **3.** Verify due diligence was refreshed at contract renewal or after a material vendor incident.	• Vendor due diligence questionnaire/report • Renewal/refresh schedule and evidence	**M**
Contract lacks AI-specific terms (model change notice, liability for hallucination/error, audit rights, data use restrictions, sub-processor disclosure)	*ISO 42001 Annex A.10; EU AI Act Art. 25*	**1.** Review the vendor contract/DPA for explicit AI clauses: advance notice of material model changes, restrictions on using customer data for vendor model training, audit/inspection rights, and liability allocation for erroneous agent actions. **2.** Flag contracts silent on these terms, particularly for Tier 1 agents. **3.** Confirm clear delineation of provider vs. deployer obligations under applicable AI regulation, where the organization is a deployer of a third-party model.	• Vendor contract / DPA / AI addendum • Provider-deployer responsibility matrix	**M**
Silent/unmanaged model version changes ('model drift' from vendor-side updates) impact agent behavior without re-testing	*ISO 42001 Annex A.6; NIST MANAGE 2.3; EU AI Act Art. 25*	**1.** Determine whether the agent uses a pinned model version or a vendor's 'latest' alias that can change without notice. **2.** Request evidence of monitoring for vendor model deprecation/update announcements and corresponding internal re-testing. **3.** Sample one vendor-driven model update in the audit period and trace whether regression testing occurred before/after the change took effect in production.	• Model version pinning configuration • Vendor update notifications received • Regression test results tied to a specific update	**H**
Concentration risk — material business processes depend on a single AI vendor/model with no tested fallback	*ISO 42001 Annex A.10; NIST GOVERN 6.2*	**1.** Identify processes where agent failure or vendor outage would materially disrupt operations. **2.** Confirm a documented fallback (alternate model/vendor, manual process, degraded mode) exists and has been tested, not just described. **3.** Review vendor SLA and incident history for the audit period; assess whether actual uptime/performance met contractual commitments.	• Business continuity/fallback plan for AI dependency • Fallback test results • Vendor SLA performance reports	**M**
No visibility into vendor's own AI Act/ISO 42001 conformity (GPAI obligations, certifications) where relevant	*EU AI Act Title V (GPAI); ISO 42001 (vendor's own certification status)*	**1.** For foundation model vendors, request evidence of their published model documentation / GPAI technical documentation (where the EU AI Act applies). **2.** Check whether the vendor holds or is pursuing ISO 42001 certification or equivalent, and request the certificate/scope statement if claimed. **3.** Assess whether vendor claims of conformity are substantiated or merely marketing assertions, by checking certification body/registry where possible.	• Vendor GPAI documentation / model card • ISO 42001 certificate and scope (if claimed) • Certification body verification	**M**

10. Monitoring, Logging & Incident Response

Standards basis: ISO 42001 Annex A.6, A.9 (Performance evaluation), A.10 (incident management); NIST MEASURE 2–4, MANAGE 4; EU AI Act Art. 12 (record-keeping/logging), Art. 26 (deployer monitoring obligations), Art. 73 (serious incident reporting).

Failure Mode / Risk	Standard Mapping	Audit Test Steps	Evidence to Request	Sev.
Logging insufficient to reconstruct an agent's decision/action chain (inadequate traceability)	*EU AI Act Art. 12; ISO 42001 Annex A.6*	**1.** Select a sample of agent actions (ideally including at least one anomalous or escalated case) and attempt to fully reconstruct the decision chain from logs: input received, reasoning/intermediate steps if captured, tool calls made, output produced, human review (if any). **2.** Assess whether logs capture sufficient detail to support a post-incident root-cause analysis, not just a final outcome. **3.** Confirm log retention period is sufficient for regulatory and dispute-resolution needs, and logs are tamper-evident (e.g., write-once storage, access-controlled).	• Raw log extracts for sampled transactions • Log retention and access control policy • Log integrity/tamper-evidence configuration	**H**
No post-deployment performance monitoring against defined KPIs/thresholds (accuracy, drift, error rate)	*ISO 42001 Annex A.9; NIST MEASURE 2.6, MANAGE 4.1; EU AI Act Art. 26(5)*	**1.** Request the defined performance metrics and thresholds for the agent (accuracy, hallucination rate, task success rate, latency, cost-per-action where relevant). **2.** Confirm monitoring is active (dashboard, automated alerting) rather than ad hoc/manual spot checks only. **3.** Test whether a threshold breach occurred in the audit period and trace the resulting action taken (or absence of action) as evidence the monitoring loop is actually closed.	• Monitoring dashboard screenshots/exports • Threshold/alert configuration • Evidence of action taken on a breach event	**H**
No defined AI-specific incident response / serious incident reporting process	*EU AI Act Art. 73 (serious incident reporting for high-risk systems); ISO 42001 Annex A.10; NIST MANAGE 4.2*	**1.** Confirm an incident response plan exists that explicitly addresses AI/agent-specific incident types (e.g., harmful output, unauthorized action taken, data leakage via prompt, injection-driven misuse) distinct from generic IT/security incident response. **2.** For high-risk systems under the EU AI Act, confirm awareness of and a process for the regulatory serious-incident reporting obligation and timelines. **3.** Sample any incident that occurred in the audit period (or run a tabletop if none occurred) and assess time-to-detect, time-to-contain, and whether lessons learned fed back into Section 4/5 controls.	• AI incident response plan • Incident log/register for the audit period • Regulatory reporting process documentation • Tabletop exercise results, if performed	**H**
User/customer-facing transparency about AI involvement is absent or inadequate	*EU AI Act Art. 50 (transparency obligations); ISO 42001 Annex A.9*	**1.** Sample customer/end-user-facing touchpoints where the agent interacts with or affects an individual. **2.** Confirm appropriate disclosure exists that the individual is interacting with or being affected by an AI system, per applicable transparency requirements. **3.** Test for dark patterns — e.g., disclosure technically present but not reasonably noticeable.	• UI/communication samples • Transparency policy • Usability/legibility assessment, if performed	**M**

11. Working Paper Template & Rating Scale

11.1 Per-Control Working Paper Fields

Document each tested control using a consistent structure to support QA review and committee reporting:

- Control reference (Domain + row number, e.g., 5.2)
- Control objective and standard mapping
- Test steps performed (cross-reference to this program)
- Population and sample size, with sampling rationale
- Evidence obtained (list documents/system extracts, with retention reference)
- Results / exceptions noted
- Root cause (where an exception is identified)
- Risk rating of the exception (see 11.2)
- Management response, owner, and target remediation date
- Auditor conclusion and preparer/reviewer sign-off

11.2 Exception Severity Rating

Rating	Definition	Reporting Threshold
High (H)	Control absent or ineffective for a Tier 1 risk; potential for material financial, legal, regulatory, or safety impact, or unconstrained autonomous action.	Report to Audit Committee; require remediation plan with executive sponsor and committed date.
Medium (M)	Control exists but is inconsistently applied, undocumented, or only partially effective; risk is contained but not eliminated.	Report in audit summary; track to closure via standard issue tracking.
Low (L)	Minor design gap or documentation deficiency with limited risk exposure; control largely effective.	Note in working papers; verify closure at next engagement.

12. Committee Reporting: Suggested Heat-Map Rollup

For Audit Committee or AI Governance Committee reporting, roll findings up by domain to show a single page risk posture view. Suggested format:

Domain	Open High	Open Medium	Open Low	Trend vs. Prior
Governance & Accountability	—	—	—	—
Excessive Agency & Authorization	—	—	—	—
Prompt Injection & Adversarial Exposure	—	—	—	—
Human Oversight & Escalation	—	—	—	—
Data Lineage, Quality & Privacy	—	—	—	—
Vendor, Model & Third-Party Risk	—	—	—	—
Monitoring, Logging & Incident Response	—	—	—	—

Note on Positioning this with the Committee

Present this matrix alongside a narrative cover memo. Boards respond to trend lines and tier-weighted exposure, not raw issue counts. Lead with Tier 1 agent coverage percentage (how many Tier 1 agents have been fully tested this cycle) — that is the single most defensible KPI for demonstrating audit coverage of agentic AI risk.

Appendix A — Standards Reference Quick Map

Domain	ISO/IEC 42001	NIST AI RMF	EU AI Act
Governance & Accountability	Clauses 5, 6, 7, 9	GOVERN 1–6	Art. 9, 17
Excessive Agency & Authorization	Annex A.6, A.8	MANAGE 1–2	Art. 14
Prompt Injection & Adversarial Exposure	Annex A.8	GenAI Profile (AI 600-1) §2.9	Art. 15
Human Oversight & Escalation	Annex A.6, A.9	GOVERN 3, MANAGE 2	Art. 14
Data Lineage, Quality & Privacy	Annex A.7	MAP 2, MEASURE 2	Art. 10
Vendor, Model & Third-Party Risk	Annex A.10	GOVERN 6	Art. 25, Title V
Monitoring, Logging & Incident Response	Annex A.6, A.9, A.10	MEASURE 2–4, MANAGE 4	Art. 12, 26, 73

Note: clause/article numbering reflects the published standards as of early 2026. Confirm against the current official text at engagement planning, as both ISO 42001 amendments and EU AI Act implementing/delegated acts continue to evolve.

Appendix B — Glossary

- **Agentic AI:** An AI system that plans and executes multi-step actions via tool/API calls with limited per-step human approval.
- **Excessive agency:** A condition where an AI agent holds permissions, autonomy, or tool access beyond what its task requires.
- **Prompt injection (direct/indirect):** Manipulation of an LLM's behavior via crafted input, either typed directly by a user (direct) or embedded in retrieved content such as documents or web pages (indirect).
- **Human-in-the-loop / human oversight:** Mechanisms ensuring a human can meaningfully understand, intervene in, and override an AI system's decisions before they take effect.
- **Data lineage:** The traceable record of a dataset's origin, transformations, and usage rights through its lifecycle.
- **GPAI:** General-Purpose AI model, a defined category under the EU AI Act subject to provider-level transparency and risk-management obligations.
- **Blast radius:** The scope of potential harm or disruption if an AI agent acts incorrectly, considering reversibility, scale, and speed.
- **Risk Tier (1/2/3):** A three-level classification assigned to each AI agent based on degree of autonomy, data sensitivity, and potential impact if the agent acts incorrectly. Determines the scope and frequency of audit testing applied to that agent (see Risk Tiering, Section 2.3).