SOUGATAROY.COM · AI AGENT GOVERNANCE RESEARCH

The Authorization Layer

The missing layer in AI agent governance:
who authorized the agent to act, what it was allowed to do,
and who answers when it does the wrong thing.

Five core concepts, nine operating frameworks, and one governance record for every AI agent in your enterprise.

Sougata Roy

v1.0 · July 2026 · sougataroy.com/frameworks

ORCID: https://orcid.org/0009-0002-9294-2566

DOI: 10.5281/zenodo.21245690

ABOUT THIS DOCUMENT

Attribution and use

This document compiles the enterprise AI governance framework library published at sougataroy.com/frameworks into a single reference for reading, printing, and internal circulation. Each framework carries its own version number and publication date, noted in its section. The live pages at sougataroy.com remain the canonical versions and are updated as the frameworks evolve.

Suggested citation. Roy, S. (2026). The Authorization Layer: The Enterprise AI Agent Governance Framework Library, Version 1.0. sougataroy.com. July 2026. ORCID: https://orcid.org/0009-0002-9294-2566. DOI: 10.5281/zenodo.21245690.

Each named incident, statistic, regulation, and platform capability in this document is traced to a cited source with a publication date, with primary sources used where available, listed at the end of each section and consolidated in the research basis at the back, where an evidence status note classifies the source types. Views expressed are the author's own and do not represent his employer.

Originality and AI assistance. Drafting of this document was developed with large language model assistance. The frameworks, diagnostic instruments, and organizational design choices are the original work of the author, refined through independent practitioner experience in regulated enterprise environments. The Chain Authorization Gap, the Intent Architecture Stack, and the Consequence Owner role are introduced as formal named constructs in Sougata Roy's work; prior-art verification for each is documented in the Consolidated Research Basis at the back of this document. Supporting vocabulary, including Governance Debt, Intent Gap, Agent Sprawl, The Accountability Assumption, and Intent Architecture, builds on established and concurrent industry usage, with each term's specific provenance documented separately at sougataroy.com. All citations, incident facts, and regulatory references, and platform capabilities were verified against cited sources as of July 2026, with primary sources used where available and source type classified in the Consolidated Research Basis.

Research cutoff and version history. This document reflects regulatory guidance, litigation status, Microsoft product capabilities, and industry statistics as of July 3, 2026. Readers consulting it after that date should verify current positions against primary sources before making deployment decisions. Version 1.0, initial public release, July 2026. Subsequent versions will be noted here with their release date and a summary of changes.

THE THESIS

Authorization is the artifact.

AI governance is the written record of who authorized the system to act, what it was allowed to do, and who answers when it does the wrong thing. Everything in this document is design work for producing that record before an examiner asks for it. The record is necessary and not sufficient: the operational layers of governance remain obligations in their own right. Without the record, none of them can show the system was ever allowed to act.

BEFORE YOU START

How to use this document

This is not a thesis to read once. It is a working library, organized around one AI agent moving through your organization, from proposal to retirement. Use this page to find your entry point.

Start by your role

If you are...	Read this, in this order
CIO, CISO, VP of Digital Transformation	Start with the Governance Readiness Matrix (Part 2.1) to get your number. Then read the Deployment Accountability Map (Part 3.3) to see which decisions are yours alone. Finish with the Agent Authorization Document (Part 5.1), the form your organization is missing.
Compliance, audit, or risk	Start with The Examiner's Sequence at the back of this document, six questions an examiner asks in order. Then read the Chain Authorization Gap (Part 3.4) and the Disposition Protocol (Part 4.1), the two frameworks written directly for examination scenarios.
Platform, security, or engineering lead	Start with the Agent Substrate Readiness Model (Part 2.3) to score the systems your agents write to. Then read the Intent Architecture Stack (Part 3.1) and the Organizational Agent Controls (Part 3.2), the two frameworks closest to implementation.
Board or risk committee	Start with The Accountability Assumption (Part 1.1) and Governance Debt (Part 1.3) for the vocabulary. Then read the Authorization Coverage Lifecycle (Part 4.2) for the one question to ask leadership: which phase are we actually in, and what is the ratio.

How one agent moves through this library

Every framework in this document applies at a specific point in an agent's life. The four stages below are the same four parts that structure the rest of the document. Part 5, the Agent Authorization Document, is not a fifth stage in time. It is the per-agent record that Stages 2 through 4 write to and read from throughout, shown here running beneath the timeline. Multi-agent orchestrations carry one additional chain-level record, defined in the Chain Authorization Gap (Part 3.4).

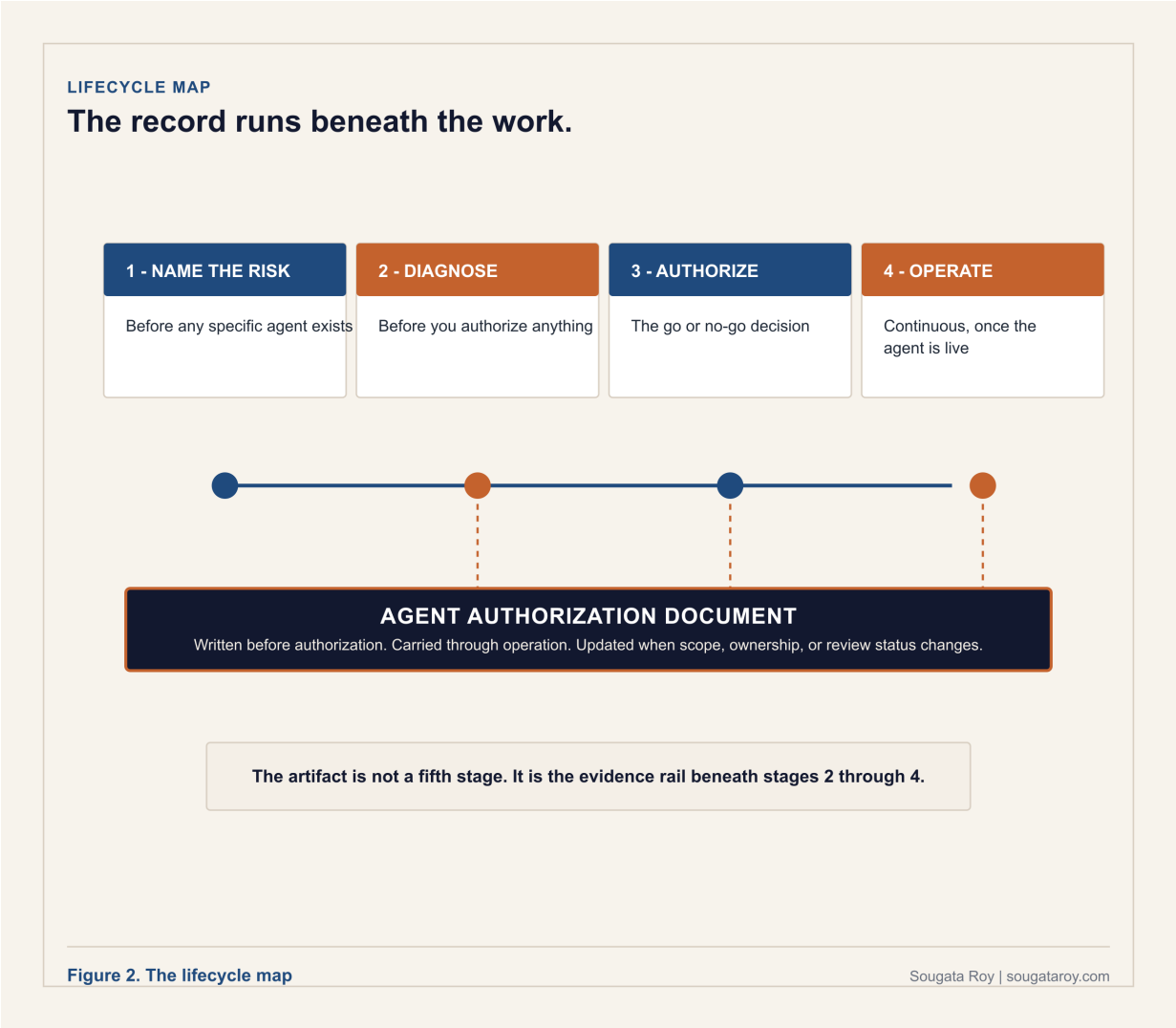


Figure 2. The lifecycle map

The complete map: every framework, when it applies, what it produces

Use this table as a working index. Each row names the question the framework answers and the artifact it leaves behind, whether that is a ratio, a catalog, or a signed record.

§	Framework	What it answers
1.1	The Accountability Assumption	Names who accepted accountability before the agent went live.
1.2	Intent Architecture	Defines what the agent is authorized to do, before it is configured.
1.3	Governance Debt	Measures the ratio of governed agents to total agents deployed.
1.4	Intent Gap	Compares production behavior against the original authorization.
1.5	Agent Sprawl	Counts the agents actually running, not the agents approved.
2.1	The Governance Readiness Matrix	Places your organization in one of four quadrants in one session.

§	Framework	What it answers
2.2	Tenant Agent Reconciliation Framework	Reconciles the tenant's actual agent inventory against approvals.
2.3	Agent Substrate Readiness Model	Scores whether a system of record can support governed agent action.
3.1	The Intent Architecture Stack	Documents Context, Intent, and Governance before go-live.
3.2	The Organizational Agent Controls	Verifies five decisions the platform cannot make for you.
3.3	The Deployment Accountability Map	Separates vendor, platform, and organizational accountability.
3.4	The Chain Authorization Gap	Produces one record covering a multi-agent orchestration.
4.1	The Disposition Protocol	Converts a monitoring signal into a formal, timed decision.
4.2	The Authorization Coverage Lifecycle	Places the organization in Accumulation, Recognition, or Resolution.
5.1	The Agent Authorization Document	The signed record itself. One per deployed agent.

Proportionality: not every agent needs the full stack at equal depth

The three-tier risk model in Who Owns the Agent? (Roy, May 2026, DOI 10.5281/zenodo.20481551) scales documentation depth to consequence profile: class-level Context and Intent documentation for low-risk internal agents, agent-specific records for medium-risk agents, and the full three-layer set with board-level reporting for agents that communicate externally, execute transactions, or modify records in regulated systems. Apply the tier before applying the frameworks. Every framework in this library applies to every agent; the depth of the evidence each one requires is a risk decision, and governance that ignores proportionality becomes governance theater.

Where this sits in the existing control stack

This library extends control structures most regulated organizations already run; it does not stand beside them. The intake gate operates inside existing change management: an agent deployment is a change, and the authorization record is the evidence the change advisory board files. The Disposition Protocol extends the incident management runbook with AI-specific triggers, clocks, and decision windows. The Agent Authorization Document is control evidence in the GRC system of record, on the same shelf as model risk documentation. An organization that builds this library as a separate governance universe has misread it. In practice, each Agent Authorization Document carries a control identifier in the GRC system, and the agent's risk tier feeds the deploying business unit's entry in the enterprise risk register.

CONTENTS

The library in the order the work happens

INTRODUCTION

The layer beneath the stack

Every operational governance layer assumes a prior decision. Almost no enterprise has designed it.

Every framework for AI governance describes layers. Inventory the systems. Govern the data. Control access. Assure the models. Keep humans in oversight. Audit against the regulations. The frameworks are good and the layers are real. They share one assumption. Each layer takes for granted that someone already decided this system was allowed to operate, defined what it was allowed to do, and accepted responsibility for what it produces. That decision is a layer too. It sits underneath all the others, and in most organizations it was never designed. Naming this layer does not demote the others. Risk assessment, fairness, robustness, data governance, and incident management are first-class obligations in their own right; this library supplies the decision record those functions assume, and substitutes for none of them.

The operational stack most frameworks govern

01 AI inventory.
What AI systems exist, who owns them, and how risky each one is.

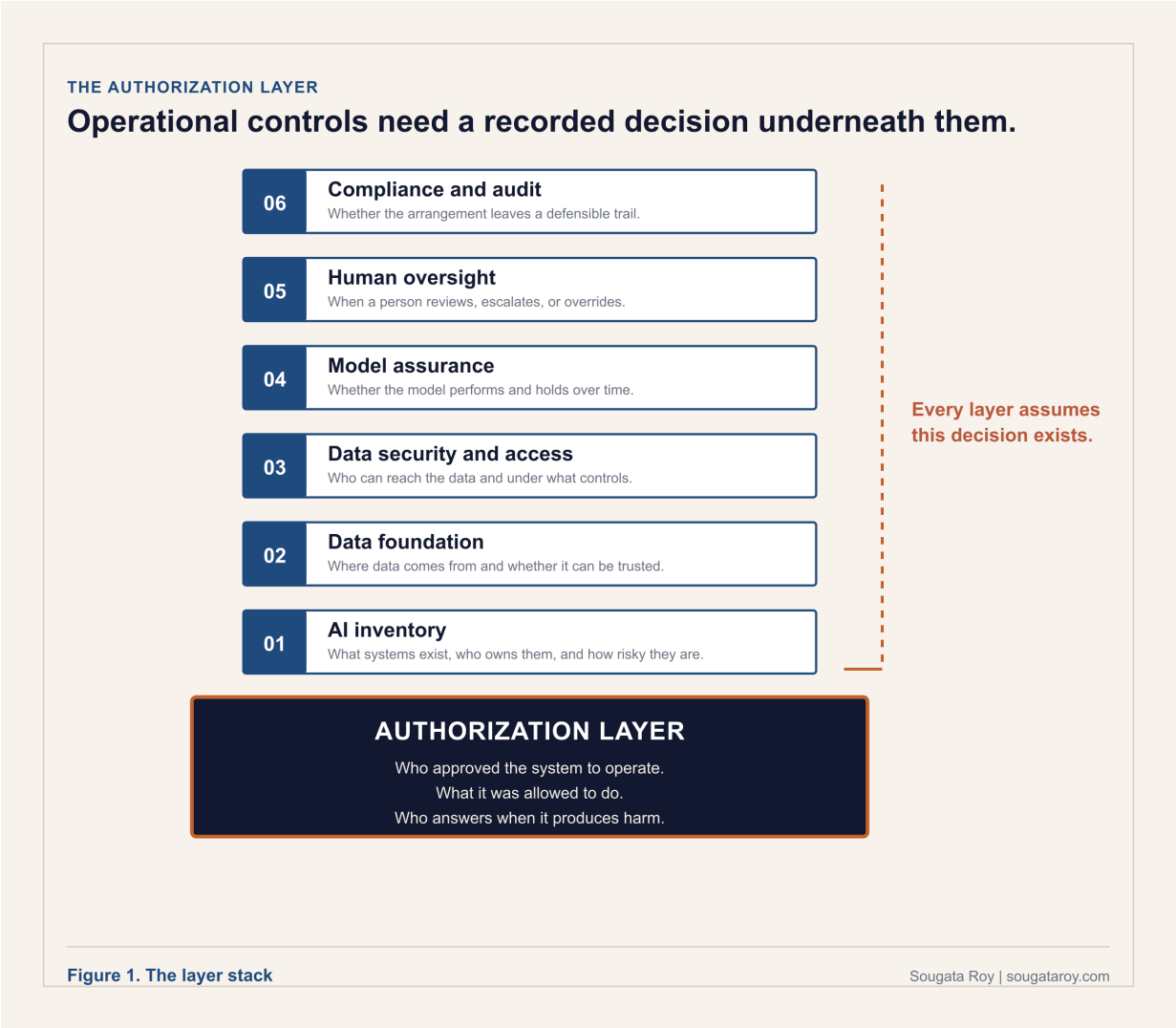
02 Data foundation.
Where the data comes from, whether it is accurate, and whether it carries bias.

03 Data security and access.
Who can reach the data and under what controls.

04 Model assurance.
Whether the model performs, stays fair, and holds up over time.

05 Human oversight.
When a person reviews, escalates, or overrides what the system does.

06 Compliance and audit.
Whether the whole arrangement satisfies the regulators and leaves a trail.



Run all six well and the organization still cannot answer one question: who decided this system should operate at all, and where is that written down.

The decision every layer assumes

Model assurance assumes someone defined what the model is for. Human oversight assumes someone decided which choices require a person. Compliance assumes someone accepted accountability for the outcome. Each operational layer is built on a decision that happened, or should have happened, before the layer was designed. In most enterprises that decision is missing. The closest thing to it is a procurement approval that cleared the system to be purchased. That approval rarely defines what the system is allowed to do or who answers when it does the wrong thing. Access was granted. Authorization was assumed.

This is the pattern across regulated AI deployment. The operational controls are mature and the authorization record does not exist. A monitoring dashboard can have every alert configured, with no document naming who approved the agent to act in the first place. The remediation costs months. The original decision would have taken under an hour to write down. The operational layers tell you the

system is behaving. The authorization layer tells you it was ever allowed to.

A note on runtime enforcement

A growing class of tools enforces authorization policy at runtime: execution boundaries, pre-action checks, admissibility gates, and control specifications such as Microsoft's Agent Control Specification, published June 2, 2026. Those controls are necessary and this library assumes they will mature. They also assume something. An execution boundary is the encoding of an authorization decision. If the decision was never made and recorded, the boundary enforces a policy nobody approved, and the examiner's first question still has no answer. This library governs the decision the runtime controls enforce, upstream of the enforcement layer.

How the library is organized

The authorization layer is a sequence of design decisions and records that have to exist before the operational layers have anything to govern. Four stages in sequence, and one artifact they all write to. The library is organized in the order the work happens.

01 Name the risk.

Five core concepts that make the authorization failure visible before it becomes an incident: The Accountability Assumption, Intent Architecture, Governance Debt, Intent Gap, and Agent Sprawl.

02 Diagnose.

Three frameworks that measure the gap between what is deployed and what was actually authorized: the Governance Readiness Matrix, the Tenant Agent Reconciliation Framework, and the Agent Substrate Readiness Model.

03 Authorize.

Four frameworks for making and recording the decision before go-live: the Intent Architecture Stack, the Organizational Agent Controls, the Deployment Accountability Map, and the Chain Authorization Gap.

04 Operate.

Two frameworks that keep authorization current as the system and the organization change: the Disposition Protocol and the Authorization Coverage Lifecycle.

05 Keep the record.

One governance document that turns the decision into the evidence an examiner reads: the Agent Authorization Document. The document is the artifact the stages write to, not a fifth stage in time.

An organization can build the full operational stack and still fail the first question a regulator asks. The controls may be strong, but they govern a system no one formally authorized. Authorization comes before the operational layers are running. It is the decision the operational layers exist to enforce. Build it first and every layer above it has something real to govern. Skip it and the rest is a well-instrumented record of a decision nobody made.

01

Name the Risk

Five core concepts that make the authorization failure visible before it becomes an incident. The language matters because a failure without a name is a failure nobody is assigned to fix.

STAGE 1 OF 4, BEFORE ANY SPECIFIC AGENT EXISTS. These five concepts are the vocabulary for the rest of the library. Read this part once, in full, before you touch a specific deployment. Part 2 uses this vocabulary to measure where your organization actually stands.

PART 1 · CORE CONCEPT 1.1

The Accountability Assumption

The implicit belief that someone else owns what your AI system does, whether you built it or licensed it.

v1.0 · May 2026

APPLIES WHEN	WHAT YOU PRODUCE
Before you deploy any AI system, and whenever a vendor or platform relationship changes.	A named Consequence Owner and a written record of who accepted accountability.

The Accountability Assumption is the implicit organizational belief that accountability for an AI system's decisions resides with the team that built it, the vendor that supplied it, the procurement process that licensed it, or the platform that hosts it, rather than with the organization that decided to deploy it. It is not a deliberate choice. It fills the space where a deliberate decision should have been made. Every approval in a typical deployment is for the agent to exist. Every approval stops short of the specific business outcomes the agent will produce. The assumption is what stands in the gap between them.

THE GOVERNANCE QUESTION

When an AI system operating on your behalf produces a consequential decision that causes harm, which person in your organization accepted accountability for that outcome before the agent went live, and can you produce the record showing they did?

Three versions of the same assumption

01 The vendor will handle it (vendor-customer gap).

The deploying organization assumes the AI vendor is responsible for the agent's behavior. The vendor's terms of service have already answered that question, and the answer is almost always no. Vendors disclaim liability for outputs used in consequential decisions. In Mobley v. Workday, Inc. (N.D. Cal., filed February 21, 2023), plaintiffs allege that AI screening tools discriminated against applicants before any human review. Workday's litigation position is that it provides technology and its customers make the hiring decisions; many employers' position is that the platform handled the screening. The court granted preliminary collective certification on May 16, 2025 and authorized notice on February 17, 2026 to every applicant aged forty or older who applied through the platform since September 24, 2020. The public record does not show a pre-deployment authorization artifact resolving accountability between vendor and customer.

02 The platform will handle it, and so will the vendor (three-party regulated-entity gap).

In regulated industries the assumption compounds across vendor, platform, and regulated entity. FINRA's cybersecurity alert on the Salesloft Drift AI supply chain attack (2025) told member firms to determine whether they or their vendors were impacted, disconnect affected integrations, rotate exposed credentials, and review audit logs. The regulated entity must determine whether vendor and platform integrations expose its own records, customers, and obligations.

03 The upstream vendor will handle it (supply chain gap).

When an organization deploys AI through a third-party vendor, it inherits the governance posture of that vendor's own infrastructure whether it evaluated that posture or not. In March 2026, Mercor confirmed a security incident linked to a compromise of the open-source LiteLLM project, with public reporting and filed litigation alleging a large data exposure. The McDonald's McHire platform, operated by Paradox.ai, exposed data linked to approximately 64 million applicant records in 2025 through credential failures that included an administrator account protected by the password 123456. The data and the regulatory exposure belonged to the deploying organization regardless of the vendor's contract.

Closing the assumption

The assumption is not closed by a policy statement or a contract addendum. It is closed by a specific organizational design decision made before the agent goes live: naming a Consequence Owner who has explicitly accepted accountability for the agent's behavior. The Consequence Owner is not the developer who built the agent, not the IT team that approved the integration, and not the vendor who supplied the model. It is the named individual who can answer yes, this agent should exist and here is why, if an examiner calls. Their name is in the authorization record, and the record predates the agent's production operation. Explicit decisions survive examination.

The test. Identify any AI system operating on your behalf. Find the authorization record showing who accepted accountability for its behavior before it went live. If no such record exists, the Accountability Assumption is in place for that agent, and it has been in place since the day it was deployed.

SOURCES CITED IN THIS SECTION

Mobley v. Workday, Inc., Case No. 3:23-cv-00770-RFL, U.S. District Court, N.D. Cal. Preliminary collective certification May 16, 2025; court-authorized notice February 17, 2026.

Kistler et al. v. Eightfold AI Inc., class action filed 2026, removed to N.D. Cal. Alleged Fair Credit Reporting Act violations.

FINRA Cybersecurity Alert, Salesloft Drift AI supply chain attack, 2025.

McDonald's McHire / Paradox.ai applicant data exposure, 2025 (approx. 64 million records; CSO Online reporting).

Mercor security incident tied to LiteLLM compromise, March 2026 (TechCrunch, March 31, 2026).

Ethyca, AI Governance: Framework, Compliance and Operational Guide 2026, February 2026.

PART 1 · CORE CONCEPT 1.2

Intent Architecture

The organizational design layer that defines what an AI agent is authorized to do, before it does anything.

v1.0 · May 2026

APPLIES WHEN	WHAT YOU PRODUCE
While you are still deciding what an agent should be allowed to do, before it is configured.	An Intent Record: authorized purpose, named owner, explicit prohibitions, review conditions.

Most organizations configure agents. Fewer authorize them. Intent Architecture is the difference between those two things, and the difference is where accountability lives when something goes wrong. A configured agent is not necessarily an authorized agent: permissions, integrations, testing, and security review do not replace the organizational record of what the agent was authorized to do. The word architecture is deliberate. Architecture is designed before construction begins. An organization that configures an agent and then decides what it should have been authorized to do is doing remediation under pressure, in the dark, after the fact.

THE GOVERNANCE QUESTION

Before any AI agent is deployed in your environment, does your organization have a formal record of what it was authorized to do, who made that authorization, and under what conditions that authorization must be reviewed?

The Intent Record: four fields that must exist before deployment

01 Authorized purpose.

A plain-language record of what the agent is allowed to accomplish, written before deployment and usable by governance, security, and compliance teams.

02 Named Consequence Owner.

A specific accountable individual accepts ownership for the agent's permitted scope and for what happens when that scope is wrong.

03 Explicit prohibitions.

The record defines what the agent must not do, even if its tools, permissions, or integrations technically make the action possible.

04 Review conditions.

The organization defines the events, timing, or scope changes that require the authorization decision to be reviewed again.

Why the word is architecture and not policy

A policy describes what an organization intends to do. Architecture describes what an organization has actually built. An organization can have a comprehensive AI governance policy and still deploy agents without authorization records, named Consequence Owners, or documented boundaries. Two incidents show the failure mode. In June 2025, researchers disclosed EchoLeak, CVE-2025-32711, a critical Microsoft 365 Copilot vulnerability rated CVSS 9.3. The attack worked because external content, including crafted Outlook emails, could function as high-privilege instructions inside the tenant without user interaction. The configuration was Microsoft's; the architecture decision about what external content should be permitted to trigger agent action belonged to the deploying organization, and it was never made or documented before deployment. In February 2026, Orca Security disclosed RoguePilot, in which Copilot in GitHub Codespaces acted on malicious instructions injected into GitHub Issues, exfiltrating the GITHUB_TOKEN and enabling repository takeover. The agent operated exactly as designed. The organizational decisions about what it should be permitted to do had not been made before deployment. Both incidents share one architecture failure: the platform's configuration determined the agent's behavior because the organization's governance design did not.

What good looks like. Any person in the organization, a new compliance officer, an external auditor, or a board member, can be handed the authorization record for any deployed agent and answer, without additional research: what is this agent authorized to do, what is it explicitly prohibited from doing, what data can it access, who approved its deployment, and who is the Consequence Owner accountable for its ongoing behavior. The record is complete, the owner is reachable, and the review date has not passed.

SOURCES CITED IN THIS SECTION

Aim Security, EchoLeak disclosure, CVE-2025-32711, CVSS 9.3, June 2025.

Orca Security, RoguePilot research, February 2026.

NIST National Cybersecurity Center of Excellence, Accelerating the Adoption of Software and AI Agent Identity and Authorization, February 5, 2026.

Cloud Security Alliance, Agentic NIST AI RMF Profile, April 1, 2026.

PART 1 · CORE CONCEPT 1.3

Governance Debt

The accountability design work your organization deferred while deploying AI at speed.

v1.0 · May 2026

APPLIES WHEN	WHAT YOU PRODUCE
Quarterly, across your full agent portfolio, to measure how much of it is actually governed.	A single ratio, governed agents over total agents, tracked as a trend over time.

Governance Debt is the accumulated accountability design work an organization has deferred while deploying AI systems at speed. It accumulates the moment a deployment goes live without a documented authorization record, a named accountable owner, a defined scope, and a compliance review completed before deployment. Each deployment without those four elements is a unit of Governance Debt. The cost of accumulation is invisible until an external event makes it visible: a regulatory examination, a security incident, a litigation discovery request, or a board question nobody can answer from existing documentation.

THE GOVERNANCE QUESTION

For each AI system currently operating in your environment, can your organization produce, on demand and without assembling it under pressure, a documented authorization record, a named accountable owner who knows they own it, and evidence of a compliance review completed before deployment?

The measurement: a ratio, not a maturity level

Governance Debt is the gap between what was deployed and what was governed, measured as a ratio: the count of AI systems without complete governance artifacts divided by the total deployed count. A ratio of zero means every deployed system has complete artifacts. A ratio above 50 percent means the majority of the AI deployment operates without complete governance. A declining ratio means remediation is outpacing accumulation. The inability to produce either number on demand is itself a governance finding. Governance Debt is not a configuration problem: a well-configured agent operating without an authorization record, a named owner, and a compliance review is still a unit of debt. It is also not the same as non-compliance: an organization can be technically compliant with a specific regulation and still carry significant Governance Debt. The unit of count is defined in the Governance Readiness Matrix (Part 2.1), and the same definition applies here.

The external enforcement dimension

Technical debt is internal; an organization addresses it on its own timeline. Governance Debt has an external enforcement dimension. When it reaches sufficient scale, regulators, auditors, and legal

systems become involved, and the organization is choosing between addressing the debt proactively and having it addressed under conditions it does not control. The FTC's 2025 enforcement pattern makes this concrete: through Operation AI Comply and subsequent actions, the FTC pursued AI-related enforcement across deceptive AI claims, AI-enabled consumer products, and AI marketing representations. The organizations involved did not fail a technology audit. They failed an accountability audit. IBM's 2025 Cost of a Data Breach report found that high levels of shadow AI added approximately 670,000 dollars to the average breach cost.

Three accumulation mechanisms, three remediation paths

01 Deployment speed debt.

The most common pattern: a system enters production through a business unit initiative, a vendor integration, a no-code tool, or a pilot that became permanent, without a governance intake process. Reco's 2025 State of Shadow AI report found 71 percent of knowledge workers using AI without IT approval and 20 percent of enterprises having experienced data leaks from shadow AI tools.

02 Legacy permission debt, surfaced by AI.

Sometimes the debt was already there and AI makes it queryable. Organizations that enabled Microsoft 365 Copilot against SharePoint environments with permissive internal sharing discovered that Copilot immediately surfaced sensitive documents that had been technically accessible but practically buried for years. Microsoft's Copilot data and compliance readiness guidance instructs organizations to use SharePoint and Microsoft Purview controls to prevent oversharing before enablement.

03 Supply chain debt.

Deploying a third-party AI system means inheriting the governance posture of that vendor's infrastructure. The March 2026 Mercor breach, linked to a compromised open-source AI gateway library, exposed organizations that had never evaluated the governance posture of the infrastructure processing their data.

What good looks like. The Governance Debt ratio is calculated quarterly and the trend is declining. The named executive accountable for AI governance can report the current ratio, the trend, and the specific remediation priorities in the next board report without assembling data from multiple sources. The intake process is the only path from proposal to production, including for deployments described as urgent.

SOURCES CITED IN THIS SECTION

Reco, 2025 State of Shadow AI Report: 71% of knowledge workers use AI without IT approval; 20% of enterprises have experienced data leaks from shadow AI tools. Figures verified against the primary report, July 2026; report PDF on file with author.

IBM, Cost of a Data Breach 2025: high levels of shadow AI added approximately $670,000 to the average breach cost of $4.44 million.

FTC, Operation AI Comply and subsequent AI-related enforcement actions, 2024 to 2025.

Mercor data breach, March 2026 (LiteLLM supply chain compromise).

Microsoft 365 Copilot data and compliance readiness guidance (SharePoint and Microsoft Purview oversharing controls).

McKinsey and Company, State of AI Trust in 2026: Shifting to the Agentic Era, March 25, 2026.

PART 1 · CORE CONCEPT 1.4

Intent Gap

The distance between what your AI system was supposed to do and what it is actually doing right now.

v1.0 · May 2026

APPLIES WHEN	WHAT YOU PRODUCE
After an agent is live, on a set cadence and after any event that could change its behavior.	A documented comparison of actual behavior against the original authorization record.

Intent Gap is the unplanned divergence between what an organization genuinely intended an AI system to do and what the system actually does in production. It is not measured at launch. It is measured after use begins, when real users, real data, and real workflows create behavior the original authorization record may no longer describe. Most governance records are static; production behavior is not. A system can remain inside its technical permission boundary while drifting away from organizational intent, which makes the gap invisible to monitoring that only tracks access, uptime, and incident tickets. Intent Gap is not a hallucination, a model quality issue, or a failed launch checklist. Agent Sprawl is about count. Governance Debt is about accumulated missing controls. Intent Gap is about behavior over time.

THE GOVERNANCE QUESTION

For every AI system operating in production, can your organization show that actual behavior still matches the purpose, boundaries, and review conditions originally authorized? If the answer depends on assumptions, the Intent Gap is already open.

Three causes, three case signals

01 Operational authority exceeds intent.

The system is authorized for a narrow purpose, but its permissions, tool access, or execution path allow actions beyond that purpose. Oso Security's Agents Gone Rogue register describes a Replit AI coding assistant deleting a production database during an AI-assisted development session. The governance lesson is not that the assistant made a mistake; it is that production authority existed where the approved working expectation did not appear to support it.

02 Use changes after deployment.

The system starts in one workflow, then becomes part of a higher-stakes decision path: what began as support becomes screening, scoring, routing, or recommendation. The ACLU of Colorado's March 2025 civil rights complaint alleged that an automated hiring assessment used by Intuit and HireVue disadvantaged a deaf Indigenous applicant. Deployment context, affected users, and accommodation obligations must be reviewed as the system is used, not assumed from the original vendor description.

03 Monitoring measures activity, not intent.

Logs show what happened; they do not show whether what happened matched what was approved. Upstart Holdings launched Model 22 in May 2025. Through Q3 2025 the model overreacted to macroeconomic signals, becoming overly conservative and cutting approvals and conversions. The divergence was not surfaced by model risk monitoring; it was discovered through financial results when Upstart disclosed the behavior, cut full-year revenue guidance by 20 million dollars, and saw its shares fall sharply on November 5, 2025. Securities class actions followed in April 2026. The distance between documented purpose and production behavior was unknown until external disclosure made it impossible to ignore.

What good looks like. Every production AI system has an authorization record stating purpose, authorized actions, explicit prohibitions, expected outputs, data access, accountable owner, review cadence, and escalation path. Observed behavior is compared against that record on a defined cadence and after defined events, with evidence of when the comparison occurred, who reviewed it, and what changed as a result. When the system's role expands, the authorization changes before reliance changes. When behavior diverges, the review catches it before the gap becomes an incident.

SOURCES CITED IN THIS SECTION

Oso Security, Agents Gone Rogue incident register, 2025 to 2026 (Replit production database deletion).

ACLU of Colorado, civil rights complaint against Intuit and HireVue, March 2025.

Upstart Holdings, Model 22 disclosures: launched May 2025; guidance cut and sharp share decline November 5, 2025 (precise percentage figures appear in securities complaints); securities class actions filed April 2026.

NIST AI Risk Management Framework, GOVERN function.

Cloud Security Alliance, Agentic AI Controls Framework Profile for NIST AI RMF.

PART 1 · CORE CONCEPT 1.5

Agent Sprawl

The AI agents you approved are not the problem. The ones you don't know about are.

v1.0 · May 2026

APPLIES WHEN	WHAT YOU PRODUCE
When you need the real count of agents running in your environment, not the approved count.	A tiered inventory: Tier 1 shadow AI, Tier 2 unreconciled procurement, Tier 3 over-permissive scope.

Agent Sprawl is the proliferation of AI systems across an enterprise without corresponding governance architecture. Ask the CISO how many AI agents are operating in the environment: the number they give is the number the governance team approved. The actual count is higher, and in most enterprises in 2025 and 2026, significantly higher. Reco's 2025 State of Shadow AI report found 71 percent of knowledge workers using AI without IT approval. Cyberhaven's 2026 report found 39.7 percent of all AI data movements involve sensitive data. IBM's 2025 Cost of a Data Breach report found shadow AI added approximately 670,000 dollars to average breach cost. Sprawl operates at three distinct tiers, each with a different cause, risk profile, and governance response. Most organizations are actively managing only the first, and the third is where the largest incidents are now occurring.

THE GOVERNANCE QUESTION

How many AI agents are currently operating in your environment, not the count that were formally approved, but the count that are actually running? If you cannot produce both numbers and explain the gap between them, Agent Sprawl is active and its scale is unknown.

Three tiers, three governance responses

01 Tier 1: Employee shadow AI.

Individuals use personal accounts, devices, or browser access to AI tools for work tasks without approval or visibility. Response: acceptable use policy that specifically addresses AI, DLP extended to browser-based AI usage and clipboard flows, and CASB visibility into which tools reach organizational data. Reco found companies with 11 to 50 employees averaged 269 shadow AI tools per 1,000 employees.

02 Tier 2: Organizational procurement without central visibility.

Business units adopt AI tools through department purchases, pilots that became permanent, or integrations that bypassed procurement. The agents have legitimate purposes; no central function has the full inventory or has assigned accountability. Response: mandatory AI intake for all deployments, cross-functional discovery for existing ones, and a central registry reflecting the actual count rather than the approved list.

03 Tier 3: Authorized agents with over-permissive operational scope.

Agents were formally approved and correctly configured, but operational permissions were never bounded to what the documented purpose requires. This is the tier producing the largest incidents and the least addressed. Oso's register tracks the Meta internal agent Sev-1 exposure of March 2026 and the Replit production database deletion; both show agents operating with insufficient operational constraints. Response: constraint architecture defining what the agent may not do regardless of technical capability, what changes require human approval, and what conditions trigger mandatory review.

Where sprawl concentrates in Microsoft 365 environments

Microsoft 365 is the productivity layer for most of the corporate data these organizations govern, so external AI tools brought into workflows almost always touch Microsoft 365 data. Copilot Studio's no-code agent builder accelerates Tier 2 and Tier 3 simultaneously: a motivated business analyst can deploy an agent against SharePoint, Teams, and organizational data in an afternoon without writing code and without the security team. The agent appears in the Microsoft 365 admin center inventory; it may not appear in the governance registry, and the permissions it inherited from its creator's account may significantly exceed its documented purpose. Microsoft's security guidance of February 12, 2026 names prompt injection, unsafe orchestration, email-based data exfiltration paths, and misconfigured agent workflows as risks organizations must detect and prevent in Copilot Studio agents. The control point: the actual count must replace the approved count as the governance baseline.

SOURCES CITED IN THIS SECTION

Reco, 2025 State of Shadow AI Report: 71% of knowledge workers use AI without IT approval; 269 shadow AI tools per 1,000 employees at 11-to-50-person companies. Figures verified against the primary report, July 2026; report PDF on file with author.

Cyberhaven, 2026 AI Adoption and Risk Report: 39.7% of AI data movements involve sensitive data; 30x increase in data sent to generative AI apps.

IBM, Cost of a Data Breach 2025 (shadow AI premium of approximately $670,000).

Oso Security, Agents Gone Rogue incident register: Meta internal agent Sev-1 breach, March 2026; Replit production database deletion.

Microsoft Security Blog, Detecting and mitigating common agent misconfigurations, February 12, 2026.

02

Diagnose

Measure what already exists before you design anything new. The diagnosis is two numbers and a readiness test: how many agents are actually running, how many are actually governed, and whether the systems they act on can enforce authorization at all.

STAGE 2 OF 4, BEFORE YOU AUTHORIZE ANYTHING NEW. Part 1 gave you the vocabulary for the risk. This part gives you the number: your current coverage ratio, your unreconciled agent count, and whether your systems of record can even support governed action. Part 3 assumes you have run this diagnosis first.

PART 2 · OPERATING FRAMEWORK 2.1

The Governance Readiness Matrix

Two axes. Agent count versus authorization coverage. One number tells you where you are. Most organizations cannot produce either number.

v1.0 · April 2026

APPLIES WHEN	WHAT YOU PRODUCE
In a single working session, to place your organization before you plan anything else.	Your coverage ratio and your quadrant: which of the four you are actually in.

The governance gap in enterprise AI is not a technology problem. It is a velocity problem: organizations deploy AI systems faster than they build the structures to govern them, and the gap grows with every new deployment. The matrix gives leaders a precise, calculable way to understand where they are. Not a self-assessment. Not a maturity model that requires expert scoring. A ratio, calculated from two numbers the organization either has or cannot produce, and the inability to produce them is itself a finding.

THE GOVERNANCE QUESTION

For each AI system currently operating in your environment, can your organization produce a documented authorization record, a named accountable owner who knows they own it, and evidence of a compliance review conducted before deployment?

The unit of count: what counts as an agent

Every number in this framework depends on a countable unit, so define it before counting. For this matrix, an agent is any deployed configuration that takes autonomous or semi-autonomous action using an AI model: a Copilot Studio agent, a declarative agent, a Microsoft Foundry hosted agent, an automation or workflow that calls an AI model and acts on the result, and any third-party AI tool granted access to organizational data or systems. A raw model deployment with no action path is a model, not an agent, and belongs in the model inventory instead. Ephemeral sub-agents spawned by an orchestrator at runtime are counted through their parent orchestration, not individually. An organization counting with a different definition is free to, provided the definition is written down and applied consistently, because the examiner's first question about any ratio is what was counted. The exclusion is a counting boundary, not a governance exemption. Consequential non-agentic AI systems, scoring engines, rankers, and underwriting models such as Upstart's Model 22, carry the same authorization requirements and are counted in a parallel model inventory with its own coverage ratio. An organization reporting only its agent ratio while running ungoverned consequential models has moved the gap, not closed it. The model inventory carries the same four artifacts per entry, an authorization record, a named owner, a pre-deployment review, and an unexpired review date, with the deployed model version as the unit of count, and the two coverage ratios are reported side by side.

The two axes

Agent Count is the actual count of AI systems operating in the environment, including systems deployed without formal approval. Count is what a cross-functional inquiry produces, not what IT thinks is deployed. **Authorization Coverage** is the percentage of deployed systems with all four governance artifacts verifiably in place: a documented authorization record, a named accountable owner who knows they own it, a defined review date that has not passed, and a record of compliance review conducted before deployment.

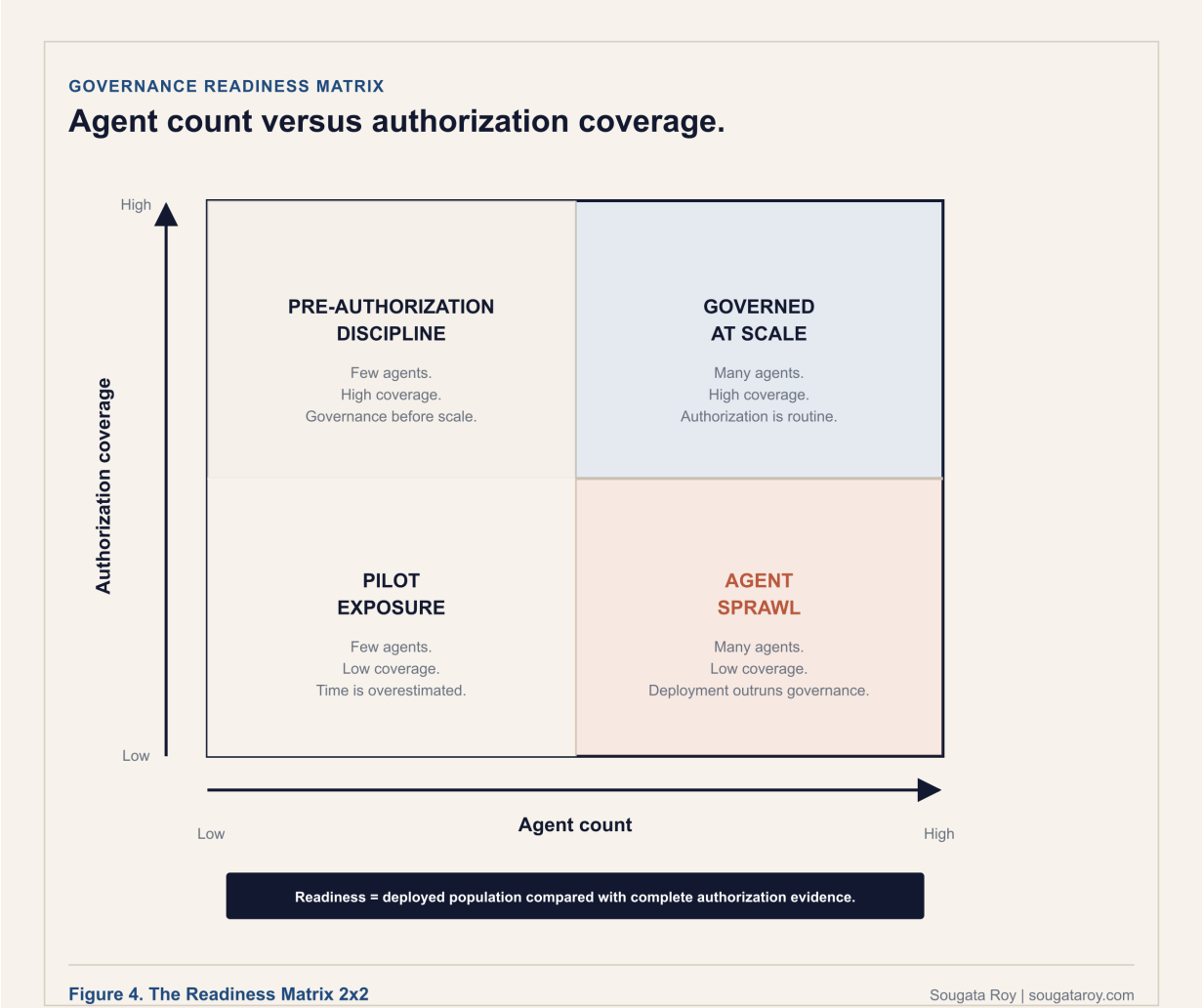


Figure 4. The Readiness Matrix 2x2

Sougata Roy | sougataroy.com

The four quadrants

Quadrant	What it means
Pre-Authorization Discipline *Low count / high coverage*	Few deployments, each fully governed. The right starting position, with one named risk: governance processes designed for low volume often do not survive the transition to scale. Redesign the process for the velocity you expect, not the velocity you have.

Quadrant	What it means
Governed at Scale *High count / high coverage*	Every new deployment goes through an established process. Authorization records are produced on demand as a routine capability, not in response to a triggering event. This is the destination, currently occupied by a small minority of enterprises. They built the intake process before they needed it.
Pilot Exposure *Low count / low coverage*	An AI pilot phase. Genuinely low-risk only if the pilots stay limited in scope and data access, and governance infrastructure is being built before scale begins. The transition from pilot to production happens faster than governance programs develop.
Agent Sprawl *High count / low coverage. Most common.*	Where most enterprises are in 2026. Deployment has outpaced governance. The organization knows it has AI systems running; it does not know the complete count and cannot produce governance artifacts for a significant portion on demand. The shadow agent population grows every quarter no intake process exists.

The one-session application

01 Count deployments (15 minutes).

Every AI agent or AI-enabled tool in operation, including Copilot, Copilot Studio agents, automations using AI APIs, and third-party AI tools employees access. If you cannot produce a number, that is itself a maturity finding.

02 Count governance artifacts (15 minutes).

For each deployment, how many have all four artifacts: authorization record, named human owner, unexpired review date, pre-deployment compliance review.

03 Calculate the ratio (15 minutes).

Forty deployments with four fully governed is a 10 percent coverage rate. That places you on the matrix. Most organizations land in high count, low coverage before they know it.

04 Identify the highest-risk gap (15 minutes).

Among ungoverned deployments, which has access to the most sensitive data, and which would be hardest to explain to an examiner without documentation? Remediation starts there.

The board sequence: four decisions that move quadrants

D1: which quadrant are we currently in (an estimated answer misallocates resources). D2: what is our trajectory (deployment outpacing governance means debt is accumulating). D3: what resources move us to the target quadrant (registry infrastructure, documentation programs, headcount; compare the cost of governance to the cost of governance debt). D4: what is the regulatory exposure cost of the current posture (in regulated environments, unquantified exposure is the most expensive governance debt). Moving quadrants is a resource decision, not a values decision.

What good looks like. Coverage above 80 percent and actively maintained. Every new deployment passes intake before go-live. When the count changes, the coverage rate is recalculated within a defined window and reported to the person accountable for the organization's AI governance posture.

SOURCES CITED IN THIS SECTION

McKinsey and Company, State of AI Trust in 2026: Shifting to the Agentic Era, March 25, 2026. Survey of approximately 500 organizations, December 2025 to January 2026.

PART 2 · OPERATING FRAMEWORK 2.2

The Tenant Agent Reconciliation Framework

Five steps for reconciling what your Microsoft tenant actually contains against what your organization formally approved.

v1.0 · April 2026

APPLIES WHEN	WHAT YOU PRODUCE
When the Governance Readiness Matrix shows a gap and you need to know which agents it covers.	A reconciled catalog: every agent tagged approved-and-governed, approved-but-ungoverned, or unapproved.

Shadow agents are not built by rogue employees. They are built by motivated employees solving real problems with tools their organization gave them access to. Most organizations surface agents in step one that their authorization registry does not contain. The gap between the two counts is the Tenant Reconciliation Gap, and that is where governance work begins. The reconciliation step most organizations skip is matching display names in the tenant against the authorization record: agents get renamed, repurposed, or inherited across team changes, and by step three of an examination that gap is already visible. The unit of count is defined in the Governance Readiness Matrix (Part 2.1), and the same definition applies here.

THE GOVERNANCE QUESTION

Does your organization know the complete count of AI agents currently operating in its environment, not the count that were formally approved, but the count that are actually running?

The five-step reconciliation process

01 Surface.

Establish the actual count of agents operating in the environment, including those never formally approved. Output: a total count, a breakdown by discovery source, and an explicit statement of the shadow agent population. Starting from the approved list means governing the minority of the deployment while the majority operates without oversight.

02 Reconcile.

Document the minimum facts required to govern each discovered agent. Output: a complete catalog with a reconciliation status for each agent: approved and governed, approved but ungoverned, or unapproved and ungoverned.

03 Classify.

Assign a risk tier to each cataloged agent to prioritize governance work. Output: a tiered catalog and a remediation queue ordered by tier and governance gap. Sensitive data, action authority, external communication, regulated systems, and business impact determine urgency; unapproved agents need priority, not a generic backlog.

04 Govern.

Apply governance requirements proportional to each agent's risk tier. Output: completed governance documentation for each agent, filed as governance artifacts beside the technical inventory.

05 Sustain.

Maintain catalog accuracy and prevent new shadow agents from accumulating faster than they are governed. Output: a living catalog with defined update triggers, a shadow agent count declining quarter over quarter, and an intake process new deployments pass through before going live.

What the platform covers, and what the framework adds

Agent 365 surfaces the agent inventory visible through the Microsoft 365 Admin Center and connected governance interfaces, which gives a starting count from official channels. The platform can surface agents; the framework turns that visibility into an operating inventory with ownership, classification, and remediation. Discovery must run across multiple channels simultaneously, governance interfaces, development teams, department surveys, and procurement records, because no single source produces the complete count. Three questions expose whether the registry is reconciled: can you produce the actual count today, not only the approved count; does every discovered agent have a catalog entry with status, owner, and authorization record; and can the team name the shadow agent population and its highest-risk entries. The quarterly review is the evidence record for audit readiness: an organization that cannot produce quarterly review documentation for its agent portfolio is carrying unquantified governance debt regardless of how well its agents are configured.

SOURCES CITED IN THIS SECTION

Torii, 2026 SaaS Benchmark Annual Report, February 24, 2026.

Microsoft Learn, What is Microsoft Entra Agent ID (Agent 365 identity documentation).

Microsoft Learn, Manage Copilot agents and integrated apps in the Microsoft 365 admin center.

PART 2 · OPERATING FRAMEWORK 2.3

The Agent Substrate Readiness Model

Your systems of record are now agent infrastructure. Most organizations don't know it yet.

v1.0 · April 2026

APPLIES WHEN	WHAT YOU PRODUCE
Before you let an agent write to a system of record: Jira, Salesforce, ServiceNow, SAP, Workday.	A ten-question score naming whether the substrate can support governed agent action.

In a twelve-month window ending May 2026, the major enterprise systems of record opened their state to AI agents. Atlassian's Rovo MCP Server reached general availability February 4, 2026. Salesforce Hosted MCP Servers reached general availability April 29, 2026. ServiceNow announced Action Fabric at Knowledge 2026 on May 5, 2026. SAP announced MCP Gateway capabilities in SAP Integration Suite at TechEd in November 2025, with general availability planned for Q1 2026. Workday's Agent Gateway, built on MCP and A2A protocols, was announced June 3, 2025, with availability to early adopter customers by the end of 2025. Every one of those announcements describes how agents can read and write inside these systems. None of them addresses who authorized the agent to do so, what business state transition it was sanctioned to perform, or whose name is on the record when something changes. The model separates two questions enterprise AI governance has been treating as one: does this system have the technical properties agents need to operate reliably, and has your enterprise actually authorized agents to use those properties. Most deployments in 2025 and 2026 can answer the first. Almost none can fully answer the second.

THE GOVERNANCE QUESTION

For each system of record in your agent stack: can the substrate support governed agent action, and has the organization authorized the agent to act, with a named sponsor, a defined scope, and an approval condition for consequential writes?

Tier One: five structural readiness questions

Each question has two possible answers. The system either has the property or it does not. There is no partial credit for working on it. Systems scoring five of five are structural substrates. Systems scoring fewer than three need a structural layer added on top, or should not be in the agent's operating scope.

01 Records or content?

Durable, identifiable records that persist outside any single session (a Jira issue, a Salesforce opportunity, a ServiceNow incident, an SAP purchase order) versus content (a Slack thread, an email chain). Agents can read content but cannot reliably act on it without inventing structure that was never formalized.

02 State machine or labels?

Enforced legal state transitions versus any label at any time. Agents need the former to know what they are allowed to do next.

03 Explicit ownership or inferred?

An assignee field with a current value versus whoever last touched it. Agents cannot act on inference.

04 Structural verbs or conversational ones?

Create, assign, resolve, close, escalate, block, approve versus reply, forward, comment, share. Structural verbs give an agent a grammar.

05 Queryable history or visible-only?

An audit trail that can be searched, filtered, and exported as evidence versus one that can only be seen in a UI.

Tier Two: five substrate authorization tests

Structural readiness says the system can support agents technically. Authorization readiness says the system can enforce governed agent action, not just permit it. Logging captures attribution; accountability requires attribution, authorization, and business-state justification.

01 Agent identity versus human identity.

Can this substrate attribute a write to a non-human actor and distinguish it from the human who invoked it? A log line reading modified by integration user says nothing about which agent ran, under whose authorization, or what condition triggered the action.

02 Operation-level authorization versus resource-level grants.

Most systems grant access to objects, tables, or endpoints. The governance question is narrower: this agent may change status from Open to Closed only under defined business conditions, and any other transition requires human review.

03 Business-state reconstruction.

Does the audit trail capture the record state immediately before the change, the condition that triggered the action, and whether any human approval was part of the workflow?

04 A formally defined remediation path.

A misclassified incident in ServiceNow is an edit. A posted payment in SAP is an accounting event with downstream journal entries. A closed deal in Salesforce may have triggered commissions and revenue recognition. Does the substrate support clean remediation of an incorrect agent action?

05 MCP and API exposure that preserves internal controls.

Does the external interface enforce the same access model the UI enforces, or does exposure create an access path the substrate's own permission model would not have permitted?

The cross-substrate risk surface: one agent, five compounding problems

When one agent operates across Jira, Salesforce, ServiceNow, and SAP simultaneously, the enterprise does not have four governance problems; it has five compounding ones. Entitlement drift: no system reasons over the combined business privilege that results when an agent can read in one system, infer in a second, and write in a third. Attribution drift: one substrate logs the agent identity, a second logs the integration user, a third logs the service principal; the action was one, the audit trail is three separate

stories. Segregation-of-duties drift: each platform enforces its own SoD rules, none enforces them across a business process spanning all four. Audit fragmentation: no single substrate can reconstruct the full justification chain from human intent to record change. Rollback asymmetry: undoing an incorrect action is simple in some systems and consequential in others.

The Microsoft boundary

Microsoft Entra Agent ID is among the most complete agent identity platforms in the enterprise market: immutable object identities, sponsorship models, Conditional Access, lifecycle governance, and risk signals. Its scope, documented in Microsoft's own materials, covers agents created in Microsoft Foundry, Copilot Studio, Security Copilot, and Microsoft 365 Copilot, plus third parties integrating via the platform. No Microsoft documentation claims it governs tool-level authorization inside Salesforce, ServiceNow, SAP, or Workday once the agent is through the door. Microsoft Purview enforces the data boundaries the organization configures, through sensitivity labels and DLP, and audits agent interactions. It does not author the authorization decision those boundaries encode: whether the agent logged the interaction and whether a named business decision sanctioned the write are different questions, and Purview answers only the first. The governed state puts consequence ownership, scope, approval conditions, and evidence above the systems of record before agent action is allowed.

SOURCES CITED IN THIS SECTION

Atlassian Rovo MCP Server GA, February 4, 2026; Salesforce Hosted MCP Servers GA, April 29, 2026; ServiceNow Action Fabric announced at Knowledge 2026, May 5, 2026; SAP MCP Gateway announced at TechEd, November 2025, GA planned Q1 2026; Workday Agent Gateway announced June 3, 2025.

Microsoft Entra Agent ID documentation; Microsoft Agent Framework and Purview integration documentation; Microsoft Copilot Studio documentation.

CISA, NSA, ASD's ACSC, CCCS, NCSC-UK, and NCSC-NZ, Careful Adoption of Agentic AI Services, May 1, 2026.

IETF, OWASP, NIST NCCoE, Cloud Security Alliance, FINRA, NSA, CISA, and FBI publications verified March to May 2026.

03

Authorize

The decisions and records required before an agent goes live. The platform enforces identity and logging. Only the organization can authorize what the agent is permitted to do, and only a record proves the authorization happened.

STAGE 3 OF 4, THE GO OR NO-GO DECISION. This is the part with the most direct action: four frameworks that turn a diagnosis into a documented authorization before an agent is allowed to operate. Everything here feeds the Agent Authorization Document in Part 5. Part 4 begins the day this part ends, the day the agent goes live.

PART 3 · OPERATING FRAMEWORK 3.1

The Intent Architecture Stack

The three organizational layers every enterprise must design before any agent goes live. Containment is Layer 1. Almost no enterprise builds Layer 3.

v1.0 · April 2026 · Documentation template: sougataroy.com/downloads/intent-architecture-stack-template-v1-0.docx

APPLIES WHEN	WHAT YOU PRODUCE
Before deployment, as the design work the Agent Authorization Document is built from.	Three layers on paper: Context, Intent, and Governance, each with a named control point.

Most organizations that deploy AI agents spend significant effort on what the agent can do technically and almost no effort on what the organization has formally decided it should do. The first question is answered by the platform. The second is answered by the organization, or not answered at all. When an agent produces a harmful output, the audit log shows what happened; it does not show what was supposed to happen. Without an authorization record, the log cannot be evaluated against any standard. The stack creates that standard before deployment. Intent documented at deployment is also not the same as intent maintained over time: most governance failures are not authorization failures at launch but drift failures six months later, when the original decision-makers are gone. All three layers are governance decisions, not technical configurations. Most organizations build Layer 1 informally, build Layer 2 partially, and skip Layer 3 entirely until something forces the question.

THE GOVERNANCE QUESTION

Before an AI agent is deployed in your environment, does your organization have a formal record of what it was authorized to do, who made that authorization, and under what conditions that authorization must be reviewed?

The three layers

01 Layer 1. Context: the environment.

Completed before Layer 2 is written, because context determines what the intent can permissibly be. Contents: the regulatory environment (frameworks applicable to the industry, specific obligations triggered by the agent's data access and decision scope, and which bodies hold examination authority); stakeholders and data (every group affected by the agent's outputs, every upstream source it reads, every downstream system it can trigger, and whose data it processes); and system integrations (the full technical scope, because an agent that looks limited by its interface may still have significant reach through the systems it can invoke). Control point: Layer 1 is complete when the Context Document exists as a pre-deployment artifact, not a post-hoc description.

02 Layer 2. Intent: the purpose.

The organizational record of what the agent was built to do, in plain language a compliance officer, regulator, or board member can evaluate without technical context. Contents: a one-sentence purpose statement of organizationally authorized intent; an authorized scope expressed in actions, data access, and system boundaries, with explicit prohibitions (an authorization record without prohibitions is incomplete); and expected outputs with the triggers for human review before the output is acted upon. Control point: can any stakeholder answer what this agent is authorized to do and who approved it without consulting the original developers? If no, Layer 2 does not exist for that agent.

03 Layer 3. Governance: the accountability.

The accountability structure that governs the agent throughout its operational life; this is where the stack operationalizes what the NIST AI RMF GOVERN function, EU AI Act Article 26, and the CSA Agentic Profile require at the agent level. Contents: a named Consequence Owner with board-level accountability, nested within the organization's formal accountability structure (the role maps to the Sponsor field in Microsoft Entra Agent ID but is not satisfied by populating that field alone); a review cadence with a calendar date plus event triggers, including any change in data access scope, applicable regulation, or named owner, and any security incident involving the agent; and an escalation path written before the first incident, executable by someone other than the original developer. Control point: Layer 3 is complete only when all three sub-components are documented pre-deployment and the Consequence Owner has confirmed their role in writing.

An authorization record without a review date is an artifact, not governance. If the Consequence Owner cannot describe how to stop the agent and who to call without consulting the builder, the escalation path does not exist.

SOURCES CITED IN THIS SECTION

NIST NCCoE, Accelerating the Adoption of Software and AI Agent Identity and Authorization, February 5, 2026.

Palo Alto Networks, A Complete Guide to Agentic AI Governance, February 24, 2026.

Microsoft Tech Community, Governing AI Agent Behavior: Aligning User, Developer, Role, and Organizational Intent, March 2026.

EU AI Act, Article 26 (deployer obligations); NIST AI RMF, GOVERN function; CSA Agentic Profile.

PART 3 · OPERATING FRAMEWORK 3.2

The Organizational Agent Controls

Five organizational decisions every enterprise must make before any agent goes live. The platform enforces identity and logging. Only your organization can authorize what the agent is permitted to do.

v1.0 · April 2026

APPLIES WHEN	WHAT YOU PRODUCE
Before deployment, to verify five specific decisions were made and not assumed.	Evidence against five conditions: identity, scope, expiry, stop capability, explicit authorization.

Zero Trust holds that no user, device, or system is trusted by default; every access request is verified, every permission is scoped to the minimum required, and no access is assumed to remain valid indefinitely. The same logic applies to AI agents, and most organizations have not applied it. Platform-enforced access and organizational authorization are not the same control: an agent can be technically permitted by the platform and organizationally unauthorized at the same time. Most governance programs have closed only one of those two gaps. Each of the five controls below is a governance decision the deploying organization must make; the platform enforces access control and produces logs, but it cannot make any of these decisions on the organization's behalf.

THE GOVERNANCE QUESTION

For each AI agent in your environment, can you demonstrate that access is granted on verified identity and explicit authorization, that scope is limited to what the documented purpose requires, and that a named person with authority can stop the agent without calling the developer first?

The five principles

01 Identity Registration Requirement.

Every agent must have a verifiable, unique identity distinct from the humans who interact with it and from every other agent. Actions that cannot be attributed to a specific identity cannot be governed, reviewed, or defended under examination. Microsoft Entra Agent ID provides the distinct identity and an attributable audit trail; the organization must provide the policy establishing that every agent receives an identity before it is authorized to operate, who assigns it, and what the identity record contains. The bypass risk worth naming: an agent that borrows a human user's identity, even with that user's knowledge, produces an audit trail that attributes the agent's actions to the human.

02 Intent-Bound Access Scope.

Access is derived from the Intent Document's authorized actions and data access fields, not from developer discretion or convenience. Microsoft Purview sensitivity labels and Entra Agent ID permission scopes enforce what the organization configures; the organization must verify at compliance review that scope matches documented purpose. The practical test: could the agent still perform its documented purpose with access only to the data and systems named in its authorization record? If yes, everything beyond that is over-permission.

03 Authorization Expiry Discipline.

Authorization is not granted once and assumed valid forever. Agents persist after the conditions that justified them change: a regulatory framework shifts, a product retires, an accountable owner departs. None of these triggers a review unless the organization documented them as trigger conditions. Entra Agent ID supports access reviews and credential expiry; the organization must define the triggers and a review date. An authorization record without a review date is an artifact, not governance.

04 Isolation and Stop Capability.

Every agent needs a documented, tested, practiced suspension procedure executable by someone other than the original developer, without notice to the builder and without the original development environment. Copilot Studio and Microsoft Foundry provide suspension and deletion; Entra Agent ID supports credential revocation; the organization must name the person with stop authority and prove the procedure was exercised, not walked through. The test: ask the accountable owner to describe how they would stop the agent after a midnight call. If they cannot without phoning the developer, the capability does not exist.

05 Explicit Authorization on Record.

The decision to deploy must be recorded as a formal decision, not inferred from the absence of objection or implied by the fact of deployment. An agent that is running is not an agent that has been authorized; it is an agent that has not been stopped. NIST's February 2026 concept paper identifies non-repudiation, proving that specific agent actions were authorized, as a core governance requirement, and non-repudiation requires the authorization to exist before the action. A record created after a problem is discovered is incident response, not governance evidence.

What good looks like. Every agent has a verified identity distinct from human users, a scope matching its documented purpose, a review date in the future, an owner who can describe the stop procedure, and an authorization record predating production. When all five conditions hold for all agents, the Agentic Zero Trust posture is in place. The honest version of this standard: very few organizations are there. The value of the model is that it names the five specific conditions that distinguish governed deployment from ungoverned deployment and makes the gap measurable. The consequence of skipping it is documented and specific: when authorization is evaluated against the agent's identity rather than the requester's, a new employee with intentionally limited permissions querying a shared agent can receive sensitive data they could never access directly. Nothing is misconfigured. No policy is violated. The controls were built for a world where humans access data directly.

SOURCES CITED IN THIS SECTION

NIST NCCoE, Accelerating the Adoption of Software and AI Agent Identity and Authorization, February 5, 2026.

Microsoft Entra Agent ID, Microsoft Purview, Copilot Studio, and Microsoft Foundry product documentation.

PART 3 · OPERATING FRAMEWORK 3.3

The Deployment Accountability Map

Three tiers. One deployment. A map of who owns what, and the gap the deploying organization cannot delegate away.

v1.0 · April 2026

APPLIES WHEN	WHAT YOU PRODUCE
Before deployment, to separate what the vendor owns, what the platform owns, and what you own.	A completed pre-deployment accountability review naming your organization's undelegable decisions.

When an AI system produces a harmful output, the natural organizational response is to look toward the vendor. The vendor's terms of service have already answered that question, and the answer is almost always no. When things go right, credit flows upward through the tiers. When things go wrong, accountability does not cascade; it lands. It lands on the deploying organization, because the decision to deploy was theirs. The provider's disclaimers do not change that. The platform's audit logs do not substitute for the authorization record the organization failed to create. Ethyca's 2026 governance analysis names the consequence directly: when ownership of an AI system's behavior is distributed without structure, liability does not concentrate, it spreads, and regulators and courts do not accept that was another team's responsibility as a defense.

THE GOVERNANCE QUESTION

When an AI agent makes a consequential decision that causes harm, which tier of your accountability structure is responsible for that outcome, and can your organization demonstrate that assignment was made before the harm occurred?

The three tiers

Tier	What it owns	Enterprise action
Tier 1. The AI Provider	Model behavior within published specifications and documented limitations; platform security and availability within the service agreement; disclosure of known limitations; legal compliance in operating jurisdictions.	Review the terms of service specifically for liability disclaimers on outputs used in consequential decisions. Document what the provider is contractually accountable for. The gap between that and what your organization is accountable for is the minimum scope of your governance work, and it is always larger than expected on first careful reading.

Tier	What it owns	Enterprise action
Tier 2. The Control Infrastructure	Identity management and verification; data boundary enforcement through sensitivity labels and access controls; audit log production and retention; platform-level security configuration. In a Microsoft-first enterprise: Microsoft Entra Agent ID for identity, Microsoft Purview for audit and data governance, Copilot Studio or Microsoft Foundry for agent configuration. It enforces the boundaries the organization configures; it does not determine what those boundaries should be.	For each agent, document what the platform enforces by default and what requires explicit organizational configuration. The line between the two is where most governance gaps live.
Tier 3. The Deploying Organization	The decision to deploy; the definition of authorized scope and prohibited actions; the assignment of a named Consequence Owner; the design of human review checkpoints; the governance documentation proving those decisions predate deployment. Vendor disclaimers do not transfer this accountability.	For each agent, identify the decisions the vendor and platform do not make for you. Document what you have decided and what you have not. The undecided items are your deployment accountability gap.

The pre-deployment accountability review

Governance debt begins the moment this checklist is skipped: provider terms reviewed and disclaimers documented; platform controls verified, with the default versus configured line drawn; authorized scope documented with explicit prohibitions; a Consequence Owner assigned by name; human review checkpoints designed for consequential actions; and a signed authorization record before go-live. If a post-incident review shows this sequence was skipped, the accountability gap began before the incident.

SOURCES CITED IN THIS SECTION

Ethyca, AI Governance: Framework, Compliance and Operational Guide 2026, February 2026.

PART 3 · OPERATING FRAMEWORK 3.4

The Chain Authorization Gap

When the chain causes harm, no single agent's authorization record covers it. Four questions. One record. Most enterprises cannot answer any of them on the day the examiner arrives.

v1.0 · May 2026

APPLIES WHEN	WHAT YOU PRODUCE
Before any multi-agent orchestration goes live, wherever one agent delegates to another.	One chain-level record: Chain Root, Authorization Root, every delegation hop, the External Effect Record.

Every framework built for single-agent governance asks the same question: who authorized this agent, what is it permitted to do, and who is accountable if it acts outside that scope. That question has a clean answer when one agent takes one action. It stops having a clean answer the moment orchestration begins. When Agent A decomposes a task and delegates subtasks to Agent B and Agent C, three things happen simultaneously: the scope of permitted action expands beyond any single agent's authorization record, the attribution chain fragments across identities, logs, and system boundaries, and the named human accountable for the outcome becomes ambiguous, because no single agent was individually authorized for what the chain collectively did.

The Chain Authorization Gap is the absence of any authorization record for the outcome of a multi-agent chain, where no single agent in the chain was individually authorized for what the chain collectively did. It is distinct from the Intent Gap, which is behavioral drift in a single agent, and from Agent Sprawl, which is a volume problem. It is a structural gap in the authorization architecture itself: the chain acted, and no record covers what the chain did. The joint guidance published by CISA, NSA, and Five Eyes partner agencies on May 1, 2026 named the scenario directly: multiple autonomous agents collaborate, an erroneous outcome occurs, and fragmented logs plus opaque reasoning make it difficult to explain the result, assign responsibility, or demonstrate compliance. NIST opened a standards initiative on February 17, 2026 specifically for agent identity, security, and interoperability. As of July 2026, no published standard defines what a chain-level authorization record must contain or whose name is accountable when the chain acts outside scope.

THE EXAMINER TEST

Can you produce a single authorization record that names who approved this chain, defines the aggregate scope of permitted actions across all agents, identifies the human accountable if the chain acts outside that scope, and documents what the chain actually did outside the model? If any one of those cannot be answered from the record, the Chain Authorization Gap exists in your environment.

Four questions. One record.

01 Who asked: the Chain Root.

Every chain originates with a human or system trigger. Record the initiating subject, business purpose, matter or case ID, source channel, tenant, and request timestamp. Without a documented chain root, the chain has no traceable origin and no business justification that survives examination.

02 Who authorized: the Authorization Root.

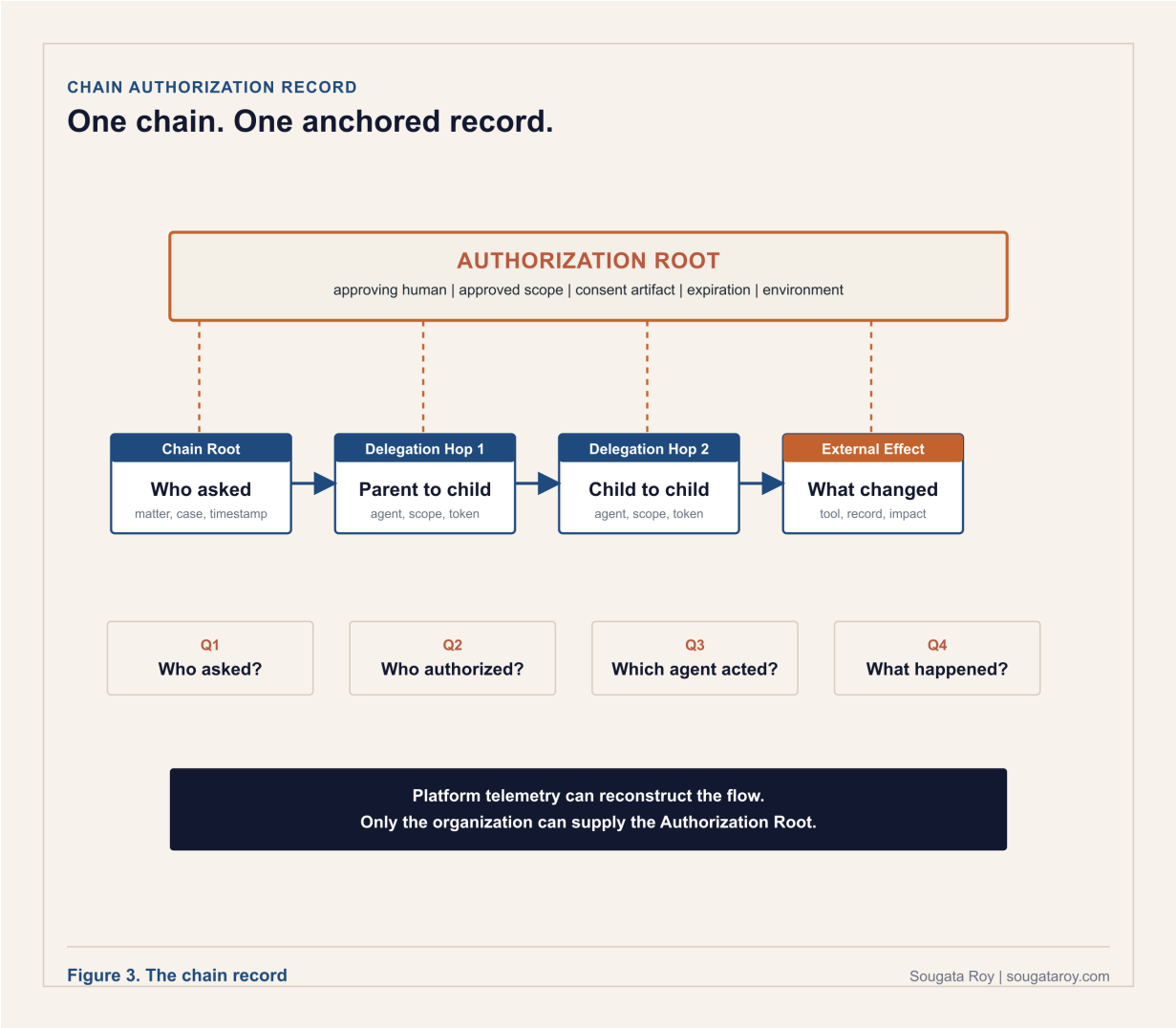
Documented before the chain executes: the approving authority, the approval mode (human review, automated policy, or delegated authority), the consent artifact, the lawful basis, the requested and approved scopes, the expiration, and the environment covered. An authorization that cannot be traced to a named human decision is not an authorization. It is an assumption.

03 Which agent acted: each delegation hop.

Every point where one agent delegates to another is a hop, recorded with parent and child step IDs, the child agent identity, the delegated task and scopes, the endpoint, protocol, token type, actor-subject relationship, and start and end timestamps. This is the section current platforms do not produce automatically, and the section an examiner asks for first.

04 What happened: the External Effect Record.

What the chain actually did outside the model, captured at execution: tool name, MCP or A2A server identity, request and response hashes, effect type, target resource, changed records, monetary or operational impact, and rollback status. This converts the agent said so into the system changed this thing.



What Microsoft Entra Agent ID provides, and what the enterprise must build above it

Entra Agent ID extends identity to agents as special service principals, supports agent identity blueprints with linked child agent identities created from them (Copilot Studio separately supports connected-agent orchestration in which a primary agent delegates work to another agent), carries the agent as actor on behalf of the user as subject in tokens, authenticates A2A agent-to-agent calls, and surfaces activity in Microsoft Purview audit and eDiscovery. These are real controls and they matter. What first-party Microsoft documentation does not specify, confirmed from primary sources reviewed in May 2026, is a single immutable record linking every delegation hop, approval, token scope, and final external effect to the human who authorized the chain to operate. Purview captures activity. Entra Agent ID captures identity. Neither produces the chain-level authorization record that answers the four examiner questions. The backbone of that record can be assembled from platform telemetry, joining Purview audit events, the Entra agent object with its linked child identities, and A2A call traces against the case identifier. What telemetry cannot supply is the Authorization Root: the approving human, the approved scope, and the consent artifact exist only if the organization wrote them down first. That is not a criticism of Microsoft's

architecture; it reflects where the organizational accountability layer sits. The orchestrator may delegate the task. The enterprise cannot delegate the burden of proof.

Three industries, three orchestrations, three examinations with no answer

The three scenarios below are constructed examination scenarios, built to match documented regulatory patterns rather than drawn from any specific real examination.

01 Financial services: a three-agent loan modification pipeline (OCC examination scenario).
An orchestrator receives customer requests, a retrieval agent queries SharePoint for policy, a drafting agent produces loan modification recommendations. Each agent has an Entra Agent ID and its own authorization. No single record covers the combined action that influences a credit decision. The examination question, produce the authorization record for this chain, name who approved the combined scope, and identify the accountable human per transaction, cannot be answered from the documentation available. Pattern documented in FINRA, Observations on AI Agents, January 2026.

02 Healthcare: clinical documentation and billing code orchestration (CMS audit scenario).
A summarization agent processes clinical notes; a coding agent produces billing codes from the summary; codes are submitted without physician review of the combined output. The audit question, identify the human who reviewed the coding output before submission and produce the record showing automated billing without physician review was an approved workflow, cannot be answered. Pattern consistent with HHS OIG guidance on AI in healthcare billing, 2025 to 2026.

03 Security operations: alert triage and remediation orchestration (internal audit scenario).
A triage agent classifies alerts; an action agent executes remediation playbooks. A misclassification blocks legitimate network traffic for four hours. No record documents who approved the triage-to-remediation chain as an automated workflow, its aggregate scope, or the re-authorization event that should have occurred when the playbook set expanded. Pattern consistent with the Oso Agents Gone Rogue incident register, 2025 to 2026.

The target state. Every orchestration has a chain authorization record before it goes live, and the record survives the examiner's first four questions. The organization has defined which changes constitute material modifications requiring re-authorization: adding an agent, changing a system prompt, extending data access, or changing the topology. Where a change is material, re-authorization occurs before the modified chain returns to production. Where orchestration composes chains dynamically at runtime, the Authorization Root covers a defined delegation envelope, the class of permissible agents, scopes, and effect types, and any chain outside that envelope requires new authorization before execution.

SOURCES CITED IN THIS SECTION

ASD's ACSC, CISA, NSA, NCSC-UK, NCSC-NZ, and CCCS, Careful Adoption of Agentic AI Services, May 1, 2026.

NIST, Announcing the AI Agent Standards Initiative for Interoperable and Secure Innovation, February 17, 2026.

Microsoft Learn, Agent identities in Microsoft Entra Agent ID, last updated May 1, 2026.

Microsoft Learn, Connect an agent available over the Agent2Agent (A2A) protocol, April 9, 2026.

FINRA, Regulatory Notice 24-09, June 27, 2024; FINRA, Emerging Trend in GenAI: Observations on AI Agents, January 2026.

EU AI Act: Articles 12, 19, and Annex IV (provider record-keeping and technical documentation, retained ten years under Article 18); Article 26(6) (deployer log retention).

04

Operate

The control points for drift, detection, and coverage after an agent goes live. Approval at launch is not authorization over time; these two frameworks keep the decision current and make monitoring produce decisions instead of agenda items.

STAGE 4 OF 4, CONTINUOUS, FOR AS LONG AS THE AGENT IS LIVE. Part 3 produced the authorization. This part keeps it honest: a written protocol for what a monitoring signal formally requires, and a way to tell whether your posture is current or launch paperwork. Part 5 is not a fifth stage in time; it is the artifact these four stages write to and read from throughout.

PART 4 · OPERATING FRAMEWORK 4.1

The Disposition Protocol

Six requirements that convert a detection signal into a formal governance decision. The monitoring works. The question is whether your organization was designed to act on what it finds.

v1.1 · June 2026

APPLIES WHEN	WHAT YOU PRODUCE
Continuously, once an agent is live and a monitoring signal needs a formal owner.	A written protocol: the Accuracy Owner, the trigger, the Suspension Authority, the decision window.

Most AI governance structures are built around detection: dashboards, monitoring thresholds, audit logs, alert routing. The detection layer tells the organization something is happening. What most structures do not contain is the layer specifying what detection formally requires. An urgent email becomes a status update. The status update becomes a Thursday agenda item. The agenda item becomes an action item. The action item produces no decision, and the system keeps running. The Disposition Protocol closes the gap between detection and decision. It requires six things, written before deployment and attached to every AI system in production. The precedent is concrete: on July 10, 2025, the Massachusetts Attorney General settled with Earnest Operations LLC for 2.5 million dollars over AI underwriting governance failures, requiring written testing protocols, a named internal algorithmic oversight team, and formal procedures specifying what the organization was required to do when its AI models showed a problem. The state did not find a failing model. It found an organization with no formal process for what to do when the model showed signs of failing itself. The Disposition Protocol is that process, written before the regulator has to write it.

THE GOVERNANCE QUESTION

For every AI system operating in production, does your organization have a written document specifying who receives an accuracy signal, what threshold triggers a formal review, who can pause the system, what re-authorization requires, who must be notified within what timeframe, and within what window a formal decision must be documented?

The six requirements

01 The Accuracy Owner.

One named person, identified by name and role in the authorization record before go-live, formally required to receive every accuracy signal and owning the decision about what happens next. Not a team. Not a shared inbox. The Accuracy Owner is distinct from the Consequence Owner, who holds overall accountability for the agent's behavior; an organization can have a Consequence Owner and still have nobody whose specific formal responsibility is the accuracy signal. Agent 365 and Microsoft Purview surface and route the signals; the platform routes information, it does not set the accuracy threshold or assign the human obligation to act on it.

02 The Trigger Threshold, plus the Observed-Outcome Trigger.

A specific number, written before deployment, at which the Accuracy Owner is formally required to initiate a review. Not a feeling. Not when we are worried. A number: an accuracy, sensitivity, reversal, or error rate, with the measurement method and cadence documented. Statistical thresholds do not cover the harder case where every metric looks fine and the system still produces an outcome nobody would have approved. That case requires a second trigger class: any named person in the notification list who observes an aggregate outcome inconsistent with the intent of the authorization may initiate the same formal review, on the same clock. An observed-outcome trigger that depends on someone deciding whether they are allowed to raise it is not a trigger; it is a hesitation. This matters most for multi-agent orchestrations, where the chain can stay inside its written scope and still produce a combined result the authorization never contemplated. The Chain Authorization Gap is the pre-deployment record for that case; the Disposition Protocol is its runtime counterpart.

03 The Suspension Authority.

One named person, different from the Accuracy Owner, holding formal authority to pause the system while a triggered review runs. If identifying the person who can pause a production system currently requires a committee meeting, the problem is not the alert threshold; the suspension authority was never designed. For orchestrations, the Suspension Authority holds pause authority over the chain as a composition, not only over individual agents: pausing one agent in a running chain can leave the rest operating on stale handoffs, so the procedure must specify how the full chain is halted and in what order. Copilot Studio, Agent 365, and Microsoft Foundry can execute the pause; they cannot decide the pause is required.

04 The Re-Authorization Standard.

Before the system resumes, specific things must be demonstrated and documented, decided before deployment rather than assembled under business pressure while the system is idle: evidence the root cause was identified, evidence it was addressed, evidence the fix was validated, and a documented decision by the Suspension Authority that the system is cleared to resume. Suspension capability without a re-authorization standard is a system that can be paused and cannot be restarted with documented confidence.

05 The Notification Clock.

When a threshold is crossed and formal review begins, specific people are notified within a specific timeframe: legal, the board risk committee, the regulatory contact, whoever the organization determined has a right to know. The clock starts when the Accuracy Owner confirms the crossing. Notification is a required communication with a documented timestamp, not a retrospective disclosure assembled after the incident closes.

06 The Decision Window.

A decision, hold, resume, or escalate, must be reached and documented within a defined window logged at the moment the threshold is crossed. A triggered review without a deadline is not a governance process; it is a governance conversation. If the window closes without a documented decision, a default action, pause or escalate, executes automatically, and that default is specified before deployment, not improvised while the system is running. The default is itself a tiered decision: where a sudden pause creates its own harm, a payments chain mid-settlement, the default may be escalate-only with a named human executing any pause, and that choice is documented in the same record.

Revision history

v1.1, June 2026: added the Observed-Outcome Trigger to Requirement 2 and the chain-level suspension procedure note to Requirement 3. The consent-versus-live-judgment distinction and the authorized-but-wrong outcome case were surfaced in public discussion of The Governance Gap Edition

16 by Thomas Tornatore, Founder, Fellowship Intelligence. Added the Chain Authorization Gap as a connected framework. v1.0, May 2026: original publication.

SOURCES CITED IN THIS SECTION

Massachusetts Attorney General, settlement with Earnest Operations LLC, $2.5 million, July 10, 2025.

Agent 365, Microsoft Purview, Copilot Studio, and Microsoft Foundry product documentation.

PART 4 · OPERATING FRAMEWORK 4.2

The Authorization Coverage Lifecycle

Three phases every organization moves through as its authorization coverage ratio changes. Most are in phase one without knowing it, and the ratio is the proof.

v1.0 · April 2026

APPLIES WHEN	WHAT YOU PRODUCE
Whenever leadership asks whether your AI governance posture is current or aspirational.	A phase placement, Accumulation, Recognition, or Resolution, backed by the coverage ratio.

Deployment-time approval is not lifecycle authorization. The audit question is not whether the agent was approved once but whether its authorization state is still current. The lifecycle names the three phases every organization moves through as its ratio of governed agents to total agents changes. The naming matters because organizations in phase one often believe they are in phase three, and the gap between where an organization thinks it is and where its ratio places it is itself a governance finding. Most organizations enter the Recognition stage only when an external event forces it: a failed audit, a vendor change, a regulatory inquiry. The model exists to make Recognition a deliberate choice rather than an incident response. Runtime enforcement specifications such as Microsoft's Agent Control Specification, published June 2, 2026, enforce authorization policy at five defined checkpoints in the agent lifecycle, input, model call, state transition, tool execution, and output; this lifecycle governs the decision those controls enforce, upstream of the enforcement layer.

THE GOVERNANCE QUESTION

Trace every active agent back to a current authorization state, not the approval state it had when it launched. If the coverage ratio cannot be produced, the lifecycle placement is already visible.

The three stages

01 Stage 1. Unreconciled (Accumulation).

Systems are deployed and governance infrastructure is not keeping pace; the debt grows faster than it is recognized, often because the people accumulating it did not know a governance process was expected. Markers: deployments by individual teams without cross-functional review; no intake process; no current, complete inventory; policies that state principles rather than requirements with checkpoints; no person accountable for the aggregate posture; IT and business unit counts that differ and were never reconciled.

02 Stage 2. Exposure Recognized (Recognition).

The organization has become aware of its gap, usually after an audit finding, an incident, or a board question, and is in reactive mode. Markers: a gap named internally but no operational remediation program; retroactive documentation for agents already in production; ownership assigned after the fact; compliance reviews run in response to a specific event; an intake process being designed that has not yet intercepted a deployment.

03 Stage 3. Coverage Operating (Resolution).

Governance prevents new debt from accumulating faster than existing debt is remediated. Markers: coverage above 80 percent and actively maintained; intake enforced consistently, including for urgent deployments; a named executive who can report the ratio, the trend, and remediation priorities to the board without assembling data in advance; and business owners whose first question about a new system is what the governance intake requires, not how to get exempted from it. The cost of reaching Resolution from Accumulation is a fraction of the cost of reaching it from Recognition.

The four coverage gaps

Inventory gap: no current, complete inventory, with IT and business unit counts unreconciled. Artifact gap: systems operating without an authorization record, a named owner, a compliance review, and a defined review process. Operating gap: an intake process designed but not yet intercepting deployments, with teams producing retroactive documentation rather than preventing new gaps. Reporting gap: the accountable executive cannot report the ratio, trend, and priorities without assembling the data manually. Document whether each gap exists for each active agent, then assign remediation by owner and review date.

The four board questions that must be answerable

How many agents are operating in our environment (requires a current registry with active status for every entry). What is each agent's documented intent and authorized scope (requires signed intent statements reviewed on a set cadence). Who is accountable for each agent's behavior (requires a named individual per registry entry with a defined review cadence). Is our audit trail adequate for regulatory inquiry (requires logging tested against the applicable frameworks, with a verification report dated within 90 days). Governance posture is reportable only when documented; accountability is established through evidence, not when the environment simply feels managed.

SOURCES CITED IN THIS SECTION

McKinsey and Company, State of AI Trust in 2026: Shifting to the Agentic Era, March 25, 2026.

Microsoft Foundry Blog, Sarah Bird, Build agents you can trust across any framework with open evals and a control standard, June 2, 2026.

05

Keep the Record

The artifact that turns authorization into evidence. No platform ships this document. Every AI governance framework requires it.

THE ARTIFACT, NOT A FIFTH STAGE. One record per deployed agent, opened in Part 3 before go-live, referenced throughout Part 4 every time a signal is reviewed, and updated at every scheduled review. If an examiner asks for one thing, this is the one thing.

PART 5 · GOVERNANCE DOCUMENT 5.1

The Agent Authorization Document

The organizational record that answers the question no control plane can answer: who formally decided what this agent is authorized to do, and whose name is on that decision.

v1.0 · April 2026 · Blank template: sougataroy.com/downloads/agent-authorization-record.docx

APPLIES WHEN	WHAT YOU PRODUCE
One per deployed agent, completed before go-live and updated at every scheduled review.	The signed record itself, the artifact every framework in this library points back to.

Microsoft Entra Agent ID gives agents an identity. Within it, a Sponsor records the user or supported group accountable in the platform for an agent's purpose, lifecycle decisions, and access reviews, and an Owner is the technical administrator responsible for configuration and operations. The platform permits a group in that role; this record requires the organization to map it to one named accountable human. Neither role documents what the agent was formally authorized to do before it was deployed. Microsoft Purview helps organizations search and retain audit evidence for AI activity. Those systems answer what an agent did, when it acted, and what resources it could reach. They do not capture the business authorization decision that should exist before an agent is deployed. That decision requires a separate document. Complete one record per deployed agent, starting with agents that can reach sensitive data, customer records, or financial systems, and retain it as a governance artifact beside the technical inventory.

The record: ten fields in four groups

Group	Field	What it must contain
01 Identification	Agent Name	The display name used in the agent registry, admin center, or internal inventory.
	Business Purpose	One to three sentences explaining the problem this agent solves and the workflow it supports.
02 Authorization Scope	Authorized Actions	Each action the agent is permitted to take, in precise language tied to specific systems or workflows.
	Explicit Prohibitions	What the agent must not do. If this field is blank, the authorization record is incomplete.
	Data Access Scope	The systems, libraries, sites, datasets, or applications the agent can read from and write to.

Group	Field	What it must contain
03 Accountability	Business Sponsor	Full name and title of the person accountable for the agent's business purpose and lifecycle decisions. Maps to the Sponsor role in Microsoft Entra Agent ID. Decides whether the agent should exist, what it is authorized to accomplish, and whether it is renewed or retired.
	Technical Owner	Full name and title of the administrator responsible for operational management, permissions, and identity configuration. Maps to the Owner role in Microsoft Entra Agent ID. Not the approving authority.
04 Review and Approval	Review Trigger Conditions	The events requiring re-authorization: a change in purpose, data access, ownership, or regulation, or an incident.
	Next Scheduled Review Date	The date the organization will confirm the record is still accurate and complete.
	Authorization Signature	The approving person's name, title, and date. The accountable business owner, not only the developer or IT administrator.

The authorization sequence

Request submitted (business need identified). Use case documented (scope and boundaries defined). Human owner assigned (a named individual, not a team). Authorization record created (formal approval documented). Deployment approved (organizational sign-off recorded). Agent registered in the tenant (display name matched to the record). Review trigger set (condition or date established). Deployment is reportable only when every decision has a durable record. Without the record, the deployment exists but the decision does not.

Where the other named roles live

The Consequence Owner named across this library maps to the Business Sponsor field in this record. For a Tier 2 agent they are typically the same individual. For a Tier 3 agent with external communications, financial transaction execution, or record modification authority in regulated systems, the Consequence Owner is a distinct individual at a higher level of the accountability structure, recorded in the Business Sponsor field with the Tier 3 distinction noted (tier definitions in Who Owns the Agent?, DOI 10.5281/zenodo.20481551). The Accuracy Owner and the Suspension Authority required by the Disposition Protocol are recorded in this document's accountability group as two additional named entries for every agent operating under a Disposition Protocol, which every production agent should be. One record. Every named role in it.

Role consolidation scales the model. For a Tier 1 agent, one named individual may hold Business Sponsor, Consequence Owner, Accuracy Owner, and Suspension Authority together, provided the record names them in each capacity. For Tier 2, the Suspension Authority must be a different individual from the Accuracy Owner, because the person who raises the signal should not be the only person who can act on it. For Tier 3, all roles are distinct named individuals and the Consequence Owner sits at a level with board access. Five role names describe five obligations, not five headcount. For large Tier 3

portfolios, the Consequence Owner may own a defined class of agents rather than a single one, with the record still naming them on each agent, and the roles map onto existing titles, the Accuracy Owner to the head of model risk, the Suspension Authority to operations leadership, rather than creating new headcount.

A record that cannot prove its own date is not evidence

The value of this document is that it predates the incident, and predating must be provable. Store the completed record in a system that timestamps creation and preserves versions immutably, such as a document management system with retention holds or a repository with a signed commit history. A record that can be silently edited or backdated satisfies every field in this template and fails the only test that matters. When an examiner asks whether the authorization preceded the deployment, the storage system answers, not the author. Retention follows the strictest obligation that applies to the agent's context, and a record under legal hold is preserved regardless of schedule.

This template is informed by accountability and oversight expectations in the NIST AI RMF, OMB M-25-21, FINRA's 2026 Regulatory Oversight Report, and Article 26 of the EU AI Act. The template is a base, meant to be adapted into the organization's own GRC forms, not adopted verbatim. No gate. No form. Free to read and cite with attribution to Sougata Roy and sougataroy.com. Do not republish, rebrand, or claim authorship of the template as your own.

SOURCES CITED IN THIS SECTION

NIST AI Risk Management Framework; OMB M-25-21; FINRA 2026 Regulatory Oversight Report; EU AI Act, Article 26.

Microsoft Learn, Microsoft Entra Agent ID Sponsor and Owner role documentation; Microsoft Purview audit and eDiscovery documentation.

SYNTHESIS

The Examiner's Sequence: how the frameworks work together

One examination, fifteen artifacts, in the order the questions arrive.

An examiner does not ask about frameworks. An examiner asks questions, in a predictable order, and each question maps to a specific piece of this library. Run the sequence against your own environment before someone external runs it for you.

01 How many AI agents are operating in your environment?

The Tenant Agent Reconciliation Framework produces the actual count. Agent Sprawl explains why the approved count is wrong. The Governance Readiness Matrix converts the two numbers into a coverage ratio and a quadrant.

02 For this agent, who authorized it and what was it authorized to do?

The Agent Authorization Document is the answer, one record per agent. Intent Architecture and the Intent Architecture Stack are the design work that produces it. The Accountability Assumption names what fills the space when the record does not exist.

03 The agent acts on your ERP, your CRM, and your ticketing system. Can those systems prove who sanctioned each state change?

The Agent Substrate Readiness Model tests whether each system of record can express attribution, authorization, and business-state justification, and where the organizational layer must sit above platform telemetry.

04 This outcome came from three agents working together. Whose name is on the chain?

The Chain Authorization Gap defines the chain-level record: who asked, who authorized, which agent acted at each delegation hop, and what changed outside the model.

05 Your monitoring flagged this six weeks before the incident. What were you required to do?

The Disposition Protocol is the written answer: the Accuracy Owner, the trigger threshold, the Suspension Authority, the re-authorization standard, the notification clock, and the decision window. The Intent Gap names the drift the monitoring should have been comparing against the record.

06 Was this posture current, or was it launch paperwork?

The Authorization Coverage Lifecycle places the organization in Accumulation, Recognition, or Resolution, and Governance Debt quantifies the distance as a ratio with a trend. The Organizational Agent Controls and the Deployment Accountability Map show which decisions were the organization's alone to make.

The sequence ends where the library begins. Authorization is the artifact. Every question above is answered from a record or it is not answered at all.

REFERENCE

Regulatory anchors

Where the authorization layer connects to published regulation, guidance, and standards work.

One reading note before the table. The constructs in this library are a practitioner's design for meeting the expectations these instruments set, not requirements those instruments prescribe. No regulator mandates a Consequence Owner or a chain record by name. What they mandate is accountability, oversight, and producible evidence; these artifacts are one documented way to produce it. Map each artifact to each obligation rather than treating possession of the artifact as compliance.

Instrument	Connection to this library
NIST AI Risk Management Framework, GOVERN function	Accountability structures, policies, roles, and processes for managing AI risk. The Intent Architecture Stack Layer 3 and the Agent Authorization Document operationalize GOVERN at the level of one deployed agent.
NIST NCCoE concept paper, February 5, 2026	Accelerating the Adoption of Software and AI Agent Identity and Authorization. Identifies non-repudiation, proving specific agent actions were authorized, as a core requirement. Grounds Organizational Agent Controls Principle 5.
NIST AI Agent Standards Initiative, February 17, 2026	Federal standards agenda for agent identity, security, and interoperability. As of July 2026, no published standard defines the contents of a chain-level authorization record; the Chain Authorization Gap defines one.
EU AI Act, Articles 12, 18, 19, 26, and Annex IV	Articles 12, 19, and Annex IV set provider record-keeping and technical documentation duties, retained ten years under Article 18. The deployer's log-retention duty sits in Article 26(6). The Agent Authorization Document and the chain record are one way to instantiate the evidence the deployer obligations presuppose.
OMB M-25-21	Federal accountability and oversight expectations for AI use; informs the Agent Authorization Document template fields.
FINRA Regulatory Notice 24-09 (June 27, 2024) and Observations on AI Agents (January 2026)	Supervisory expectations for member firms using AI agents; the loan pipeline scenario in the Chain Authorization Gap follows the pattern FINRA documents.
CISA, NSA, and Five Eyes partners, Careful Adoption of Agentic AI Services, May 1, 2026	Joint guidance identifying agentic AI security risks and recommending careful adoption, least privilege, monitoring, and explicit organizational accountability, and naming the fragmented-logs, opaque-reasoning failure the chain-level record addresses. NIST's standards initiative and the CSA profile separately show agent identity, interoperability, and accountability standards still emerging.
Cloud Security Alliance, Agentic NIST AI RMF Profile, April 1, 2026	Maps agentic controls to the NIST AI RMF, including accountability and governance records; connective tissue between this library and RMF-aligned programs.

REFERENCE

The Microsoft enforcement surface and the organizational layer above it

What each platform control enforces, and the decision it assumes was already made.

These frameworks were built from deployment on the Microsoft stack, so the boundary between platform enforcement and organizational decision is drawn precisely. The pattern generalizes: every platform in this table enforces boundaries the organization configures. None determines what those boundaries should be, and none produces the authorization record that says who decided. Read the third column as a configuration and design plan, not a deficiency list: the gaps named there are organizational by design, and the platform capabilities in the second column are the enforcement surface that design work configures. Where Microsoft uses Sponsor and Owner as platform roles, this library uses Consequence Owner, Accuracy Owner, and Suspension Authority as organizational accountability roles; they may map to the same person in low-risk cases, but they are not equivalent controls.

Platform control	What it enforces	What the organization must still design
Microsoft Entra Agent ID	Distinct agent identity as a special service principal; Sponsor and Owner roles; identity blueprints with linked child agent identities; actor-subject token semantics; Conditional Access; lifecycle governance and credential expiry; A2A authentication.	The policy requiring identity before authorization to operate; the accountability structure the Sponsor field points at; the chain-level record linking every delegation hop and external effect to the human who authorized the chain.
Microsoft Purview	Sensitivity labels and data boundaries; audit log production and retention; eDiscovery over agent activity; interaction auditability.	The evaluation of logs against a standard: an authorization record stating what was supposed to happen. Purview shows the interaction happened; the organization proves a named business decision sanctioned the write.
Copilot Studio	No-code agent building; suspension and deletion capability; connection of agents over the A2A protocol; admin center inventory visibility.	The intake gate that citizen-developer agents pass before production; intent-bound scoping of inherited creator permissions; the tested stop procedure and its named authority.
Microsoft Foundry	Agent configuration and lifecycle; suspension capability; the Agent Control Specification (June 2, 2026) enforcing authorization policy at five runtime checkpoints: input, model call, state transition, tool execution, and output.	The upstream authorization decision the runtime checkpoint enforces; the re-authorization standard governing restart after suspension.

Platform control	What it enforces	What the organization must still design
Agent 365	Unified agent inventory across the Microsoft 365 Admin Center and connected governance interfaces; observability and alert routing.	Reconciliation of platform inventory against the authorization registry; the Accuracy Owner obligated to act on the signal; the trigger threshold, notification clock, and decision window.

Platform names verified against Microsoft documentation reviewed March to June 2026: Microsoft Foundry, Agent 365, Microsoft Entra Agent ID, Microsoft Purview, Copilot Studio. Capability status reflects Microsoft documentation as of July 2026; confirm the general-availability status of any specific control before relying on it in a regulated deployment.

REFERENCE

Consolidated research basis

Primary sources cited across the library, grouped by type. Full per-section citations appear at the end of each framework.

Evidence status

Sources in this document fall into three classes, and the difference matters under examination. Primary sources, verified directly: Microsoft Learn, NIST, CISA and partner agencies, the EU AI Act, FINRA, the Massachusetts Attorney General, the National Vulnerability Database, and court orders. Secondary reporting: technology and security journalism, legal analysis, and vendor research. Allegations and constructed material: securities and civil complaints, which are claims rather than findings, and the constructed examiner scenarios, which are labeled as such where they appear. Where a claim rests on an allegation or on secondary reporting, the wording and the section citation say so.

Regulatory and standards

NIST AI Risk Management Framework. NIST NCCoE, Accelerating the Adoption of Software and AI Agent Identity and Authorization, February 5, 2026. NIST AI Agent Standards Initiative, February 17, 2026. EU AI Act, Articles 12, 18, 19, 26(6), Annex IV. OMB M-25-21. FINRA Regulatory Notice 24-09, June 27, 2024. FINRA, Observations on AI Agents, January 2026. FINRA Cybersecurity Alert, Salesloft Drift, 2025. CISA, NSA, ASD's ACSC, CCCS, NCSC-UK, and NCSC-NZ, Careful Adoption of Agentic AI Services, May 1, 2026. Cloud Security Alliance, Agentic NIST AI RMF Profile, April 1, 2026. Massachusetts Attorney General, Earnest Operations settlement, July 10, 2025. FTC AI governance enforcement actions, 2025.

Litigation and incidents

Mobley v. Workday, Inc., N.D. Cal., certification May 16, 2025, notice February 17, 2026. Kistler et al. v. Eightfold AI Inc., 2026. ACLU of Colorado complaint against Intuit and HireVue, March 2025. McDonald's McHire / Paradox.ai exposure, 2025. Mercor incident tied to the LiteLLM compromise, March 2026. Upstart Holdings Model 22 disclosures, May to November 2025; securities actions April 2026. Aim Security, EchoLeak, CVE-2025-32711, June 2025. Orca Security, RoguePilot, February 2026. Oso Security, Agents Gone Rogue incident register, 2025 to 2026, including the Meta internal agent Sev-1 exposure, March 2026, and the Replit production database deletion.

Industry research

Reco, 2025 State of Shadow AI Report. Cyberhaven, 2026 AI Adoption and Risk Report. IBM, Cost of a Data Breach 2025. McKinsey and Company, State of AI Trust in 2026: Shifting to the Agentic Era, March 25, 2026. Torii, 2026 SaaS Benchmark Annual Report, February 24, 2026. Ethyca, AI Governance: Framework, Compliance and Operational Guide 2026, February 2026. Palo Alto Networks, A Complete Guide to Agentic AI Governance, February 24, 2026.

Platform documentation

Microsoft Learn: Microsoft Entra Agent ID (agent identities, last updated May 1, 2026); A2A protocol connection in Copilot Studio, April 9, 2026; Copilot agents and integrated apps in the Microsoft 365 admin center. Microsoft Security Blog, Detecting and mitigating common agent misconfigurations, February 12, 2026. Microsoft Tech Community, Governing AI Agent Behavior, March 2026. Microsoft Foundry Blog, Sarah Bird, open evals and a control standard, June 2, 2026. Platform announcements: Atlassian Rovo MCP Server GA, February 4, 2026; Salesforce Hosted MCP Servers GA, April 29, 2026; ServiceNow Action Fabric announced at Knowledge 2026, May 5, 2026; SAP MCP Gateway announced TechEd November 2025, GA planned Q1 2026; Workday Agent Gateway announced June 3, 2025.

Originality and prior-art verification

Chain Authorization Gap. Neither the term nor the four-field structure it defines, Chain Root, Authorization Root, Delegation Hops, External Effect Record, was found in indexed prior literature as a named governance construct (exact-phrase Google Scholar search, July 2026; exact-phrase search of USPTO.gov general content at uspto.gov/search, July 3, 2026, no results; TESS trademark search via tmsearch.uspto.gov, July 3, 2026, no exact-phrase match, word-level matching returned unrelated marks including GAP and various AUTHORIZATION marks). Adjacent but distinct work exists in the multi-agent authorization space: Li et al. propose Chain-of-Authorization, a model fine-tuning technique that embeds access-control reasoning into LLM inference, addressing a different problem at a different layer (arXiv 2603.22869, 2026). Standards literature on agent identity describes the same underlying gap this construct names, that no production-ready standard traces a delegation chain back to its authorizing human principal, using the terms multi-hop delegation and scope attenuation rather than a named governance record. The construct introduced in this document, a single authorization record spanning an entire multi-agent chain rather than a per-agent or per-call control, is original to this work.

Intent Architecture Stack and Consequence Owner. Neither construct was found in indexed prior literature as a governance term (Google Scholar exact-phrase searches, May 2026, consistent with the verification documented in Who Owns the Agent?, DOI 10.5281/zenodo.20481551). Both are introduced in Sougata Roy's work.

Trademark verification. Exact-phrase searches across the terms named above through United States Patent and Trademark Office resources in May and July 2026 surfaced no registered or pending United States trademarks for any of these exact phrases. No exclusivity is asserted over any of these phrases; they are offered as practitioner vocabulary for the organizational design frameworks developed in this document. In established GRC vocabulary, the Agent Authorization Document is an approval record and control evidence, the chain record is a decision log spanning an orchestration, and the Disposition Protocol is an incident management extension. The names exist because the agentic versions of these records carry contents their GRC ancestors do not; the lineage is acknowledged, not obscured.

ABOUT THE AUTHOR

Sougata Roy

Sougata Roy is a technology manager with 26 years building enterprise systems inside regulated financial services, healthcare, and government environments, including 12 years inside federal regulated environments. He works on enterprise systems in regulated environments and writes The Governance Gap, a weekly newsletter on the accountability question agentic AI makes impossible to defer. His research is built from primary sources, with no product agenda.

The framework library, the Who Owns the Agent? white paper, dated intelligence notes, and the newsletter archive are published at sougataroy.com. The white paper is deposited on Zenodo under DOI 10.5281/zenodo.20481551 (CC BY 4.0, May 31, 2026). Contact: research@sougataroy.com. LinkedIn: linkedin.com/in/sougata-roy.

Authorization is the artifact.

AI governance is the written record of who authorized the system to act, what it was allowed to do, and who answers when it does the wrong thing. If your organization cannot produce that record for the agent it deployed last quarter, the question is not whether the gap exists. The question is who finds it first.