FRAMEWORK WHITE PAPER · VERSION 1.0

Who Owns the Agent?

The Organizational Accountability Architecture That Existing Governance Frameworks Require But Do Not Implement at the Agent Level

The Intent Architecture Stack

Sougata Roy | sougataroy.com | May 2026

ORCID: https://orcid.org/0009-0002-9294-2566
DOI: 10.5281/zenodo.20481551
Views expressed are personal. Not legal or regulatory advice.

Originality, attribution, and AI assistance. Drafting of this paper was developed with large language model assistance. The contribution is the integration of established and concurrent enterprise governance vocabulary into a coherent organizational design framework for agentic AI accountability. The three-layer Intent Architecture Stack, the three-tier risk model, the Governance Debt Maturity Model with its three stages (Accumulation, Recognition, Resolution), the Consequence Owner role construct, and the diagnostic instruments are the original work of the author, with public documentation in The Governance Gap newsletter on LinkedIn, editions 1 through 11 (March through May 2026), and at sougataroy.com framework pages from the same period. The three-layer Intent Architecture Stack, the Consequence Owner role, and the three-tier risk model are first introduced as formal named constructs in this paper. Supporting terms used throughout the paper, including Governance Debt, Agent Sprawl, and Intent Gap, appear in established or concurrent industry and academic literature; attribution and the specific application of each to agentic AI accountability are documented in the About This Paper section. All quoted material is attributed inline to its primary source.

Contents

Executive Summary

There is a meeting that is about to happen inside your organization. You may not know it is scheduled yet.

An AI agent is going to do something unexpected. It will not be a dramatic failure. It will be something quieter: an email sent to the wrong party, a document shared outside its intended scope, a decision made at 2am by a system that was technically operating within its permissions but well outside what anyone intended. The logging system will capture all of it in perfect detail. Then someone will ask the question that no logging system answers: who in this organization was responsible for what that agent just did?

If that question does not have a pre-written answer before the incident occurs, the organization faces a board conversation it is not prepared to have.

This white paper introduces the Intent Architecture Stack, a three-layer organizational design framework that answers that question before the meeting is called. The three layers are Context, Intent, and Governance. Each layer defines a specific organizational condition that must be present before an AI agent is authorized to operate in production. Together, they produce the governance architecture that existing compliance frameworks require but do not define.

Five major frameworks require that someone be accountable when an AI agent takes an action: NIST AI RMF, the EU AI Act, ISO/IEC 42001, the Cloud Security Alliance's Agentic Trust Framework, and Microsoft's Responsible AI Standard. None of them specify the organizational design that makes accountability operational at the agent level. The Intent Architecture Stack provides that design. It operationalizes NIST's GOVERN function, instantiates EU AI Act Article 26's human oversight requirement, and gives the CSA Agentic Profile's accountability register its organizational structure in Microsoft-first enterprises.

Who this paper is written for. This paper serves three readers. The CISO or CTO who needs to answer the board's accountability question before the next agentic AI deployment. The security architect who needs an organizational design framework to work against. The board member who needs to understand what governance evidence they should be asking for.

Vocabulary: five terms used throughout this paper

Intent Architecture. The organizational design layer that defines what an AI agent is authorized to do, who authorized it, and what happens when it acts outside those boundaries. Built before deployment. The Intent Architecture Stack is the operational framework that implements this concept across three layers.

Intent Gap. The unplanned divergence between what an organization genuinely intended an AI system to do and what it actually does in production. It is not a deployment failure. It is a monitoring failure that accumulates during operation when no mechanism exists to compare documented intent against actual behavior on an ongoing basis.

Governance Debt. The accumulated accountability design work an organization 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. Unlike technical debt, Governance Debt has an external enforcement dimension. When it reaches sufficient scale, regulators, auditors, and legal systems become involved.

The Accountability Assumption. The implicit organizational belief that accountability for an AI agent's decisions resides with the vendor that supplied it, the platform that hosts it, or another team, rather than with the organization that decided to deploy it. It is not a deliberate choice. It is what fills the space when no deliberate accountability assignment is made. Regulators in financial services, employment, and consumer products have emphasised that firms remain responsible for legal compliance when using third-party AI systems and that vendor contracts or terms of service do not shift that responsibility.

Agent Sprawl. The proliferation of AI agents across an enterprise without corresponding governance architecture. It operates across three distinct tiers: employee shadow AI (individuals using unsanctioned tools), organizational procurement without central visibility (business units adopting AI independently), and authorized agents with over-permissive operational scope (governed agents whose operational boundaries were never formally defined). Each tier requires a different governance response. Solving the first tier does not address the second or third.

The Intent Architecture Stack, Governance Debt Maturity Model, Consequence Owner role, and diagnostic instruments are introduced in this paper. Related vocabulary including Intent Gap, Intent Architecture, Governance Debt, and The Accountability Assumption appears in The Governance Gap newsletter (editions 1 through 11, March 20

through May 26, 2026) and at sougataroy.com/frameworks from the same period; several of these terms also appear in concurrent or prior literature in adjacent domains, as noted in the About This Paper section. Free to cite with attribution.

Section 1: The Question Nobody Has a Good Answer For

The meeting had been going for forty minutes when the board chair held up a printout and asked a simple question.

Three weeks earlier, the organization's Copilot Studio agent had sent a draft contract amendment to an external counterparty. The agent had been configured to draft for human review before sending. The human review gate was not enforced. The version it sent was not the current version. The compliance team had logs. The security team had the Purview audit record. The IT team had the agent configuration file. What none of them had was a document written before the incident that answered the chair's question: who in this organization authorized that agent to act, and how is that authorization documented?

The CISO had three answers ready. All three were technically accurate. None of them were what the chair was asking for. The chair was not asking what the agent did. The chair was asking who owned it.

> **Logging tells you what happened. It does not tell you who was responsible. Those are different questions, and only one of them survives a board meeting.**

The Statistics That Describe the Posture

Figure	Source
86% of organizations lack or do not enforce access policies for AI identities; only **16%** govern AI access to core business systems effectively.	Saviynt, 2026 CISO AI Risk Report.
83% of organizations planned to deploy agentic AI capabilities into business functions; only **29%** felt ready to do so securely.	Cisco State of AI Security 2026.
37% of CISOs cite securing AI agents as their most urgent concern, above employees' use of	Team8 CISO Village Survey, July 2025.

Figure	Source
AI tools (36%), which ranked second. This is the issue at the top of the stack.	

What the Incidents Show

The statistics above describe a posture. The cases below describe what that posture looks like when an incident converts the governance gap from a theoretical problem into an operational one. The cases below are drawn from court rulings, published company statements, regulatory filings, and security research.

Air Canada, February 2024: the ruling that rejected the chatbot-as-separate-entity defense. Air Canada argued in court that it could not be held liable for its chatbot's statements because the chatbot was, in effect, a separate entity from the company. The British Columbia Civil Resolution Tribunal rejected this argument. As the Tribunal put it in *Moffatt v. Air Canada*, 2024 BCCRT 149, "It should be obvious to Air Canada that it is responsible for all the information on its website. It makes no difference whether the information comes from a static page or a chatbot." The organization owned the agent's behavior. The ruling did not explain how that accountability should be designed into an organization before an incident. That is the gap this paper addresses.

Microsoft Copilot, February 2026: when platform configuration is not a governance record. In February 2026, Microsoft acknowledged in advisory CW1226324 that Copilot Chat had incorrectly processed some customer email messages, including content carrying sensitivity labels and subject to DLP policies. Microsoft stated that access controls remained intact and that the bug caused Copilot to behave contrary to its intended protection model rather than bypassing underlying authorization. The governance lesson does not depend on whether the controls failed technically. It depends on what happened next: the accountability question shifted immediately from what the platform was configured to do to who in the organization had explicitly authorized that AI scenario and owned the consequence. Organizations with a written authorization scope had an answer they could give to a board or an examiner before the investigation began. Organizations whose governance artifact was the platform configuration had nothing to show except how they had set up a tool. When that tool behaved unexpectedly, no document existed that named who in the organization had authorized that scenario and who was responsible for what happened next.

AWS, December 2025: governance failure with an AI tool in the loop. In December 2025, AWS suffered a prolonged service disruption that media reports linked to changes made using its internal Kiro AI coding tool; Amazon has stated that the root cause was a manual configuration mistake, not a failure of the AI system itself. Subsequent reporting indicates Amazon tightened approval requirements for AI-assisted production changes after the incident. AWS has not published a detailed public post-incident report. The governance lesson is precise: the authorization boundary existed in policy but was not enforced in practice, and no organizational design work had been done to ensure the two matched. When AI tools participate in production changes, the approval boundary is no longer a human-to-human handoff that can be assumed. It must be explicitly documented, assigned, and enforced at the organizational level before the tool is used in that context.

Upstart Holdings, May 2026: the Intent Gap priced by a securities filing. Upstart Holdings launched its Model 22 AI underwriting system in May 2025, describing it to investors as a tool that would increase loan approvals and improve risk assessment accuracy. During Q3 2025, management disclosed that the model responded to macroeconomic signals by tightening credit, reducing borrower approvals and conversion rates from 23.9 percent in Q2 to 20.6 percent in Q3. Revenue guidance was cut by twenty million dollars. The stock fell 9.71 percent on November 5, 2025. Securities class actions followed in 2026, alleging the company had not adequately disclosed the model's behavior and its impact on revenue (see Dunn v. Upstart Holdings, Inc., No. 3:26-cv-02974, United States District Court for the Northern District of California, complaint filed April 7, 2026, before Judge James Donato; the action names Upstart Holdings and individual officers as defendants and remains in the pleading stage as of May 28, 2026, per the public PACER docket via CourtListener. No ruling on the merits has been issued). The governance lesson is distinct from the securities question now before the courts: the available record suggests that Model 22's intended role was documented at launch while its live production behavior diverged in ways that were not surfaced until a quarterly earnings disclosure. Whether a monitoring mechanism existed and whether it functioned as designed are questions the litigation may answer. The observable governance failure is that the divergence between documented intent and production behavior was discovered through investor reporting rather than through internal governance.

Kistler v. Eightfold AI, January 2026: the Accountability Assumption in hiring. A class action filed against Eightfold AI in California state court in January 2026, and later removed to the Northern District of California, alleged that Eightfold AI scraped data on a

large number of workers, scored applicants on a numerical scale, and had low-scoring applicants removed from consideration before any human review occurred. Plaintiffs alleged that Eightfold operated as a de facto consumer reporting agency without FCRA disclosures, access, or dispute rights, and that employers relied on its scores without informing applicants. The complaint's central governance implication is that neither Eightfold nor the employer appears to have established clear ownership of legal responsibility for the screening decisions: the vendor and employers each left FCRA compliance and disclosure obligations unassigned. This is the Accountability Assumption in its most direct form: the implicit belief that legal responsibility for an AI system's consequential decisions resides with the other party. (Kistler et al. v. Eightfold AI Inc.: complaint filed in the Superior Court of California, County of Contra Costa, on January 20, 2026, No. C26-00214; removed by the defendant to the United States District Court for the Northern District of California on March 2, 2026, No. 3:26-cv-01768, now before Judge Yvonne Gonzalez Rogers. The defendant filed a motion to dismiss on April 20, 2026, with a hearing set for August 4, 2026; no ruling on the merits has been issued as of May 2026, per the public PACER docket via CourtListener.) A parallel pattern was reported in connection with McDonald's McHire platform, operated by Paradox.ai, which according to technology press accounts exposed applicant personal data through a credential vulnerability; no official primary disclosure from McDonald's or Paradox.ai has been confirmed. The reported incident illustrates the same vendor-governance gap: when an AI vendor handles consequential applicant data at scale, the question of who owns security review and compliance posture must be answered before deployment, not after a press account surfaces the exposure.

Auditability Is Not Accountability

Each organization in the operational incidents above had logging. Several of them had sophisticated logging. The logs were not the problem. The problem was that the logs answered a question nobody was asking in the moment, while the question everyone was asking had no pre-written answer.

Auditability and accountability are not the same thing, and confusing them is the most expensive governance mistake a regulated enterprise can make.

Auditability means the organization can reconstruct what happened. Purview captures what agents accessed and what they processed. Entra Agent ID logs authentication events. Defender for Cloud logs behavioral anomalies. These systems are designed for auditability, and in most regulated enterprises they do their job well.

Accountability is a different condition entirely. It means the organization can identify, from a document written before the incident, who authorized the agent's action scope, who owns the consequence of that action, and what organizational structure connects those two things. Norval, Cobbe, and Singh (2022) define accountability in AI systems as "answerability": the capacity to call a specific party to account, demand justification, and enforce a consequence if the justification fails.

An organization can have complete auditability and zero accountability simultaneously. The logs show every action the agent took. The board asks who authorized those actions. The answer is: it was in the configuration. That answer ends careers on a Tuesday morning, and it is entirely preventable.

Section 2: Five Frameworks, Five Different Versions of the Same Gap

Before introducing the Intent Architecture Stack, the major governance frameworks deserve a precise reading. This is not a critique. They are serious documents, and nothing in any of them prevents per-agent accountability documentation. The Intent Architecture Stack is designed to operationalize what they require. It instantiates NIST AI RMF's GOVERN function at the agent level, concretizes EU AI Act Article 26's human oversight requirement for agent-specific deployments, and gives the CSA Agentic Profile's accountability register its organizational teeth in Microsoft-first enterprises. The gap is an implementation gap, not a conceptual one. These frameworks require accountability at the organizational level. What enterprise practice still lacks is a consistent way to make that accountability operational for each deployed agent before it goes live.

Here is the enterprise absurdity buried in the current state of AI governance. An organization can be fully compliant with NIST AI RMF, fully aligned with ISO 42001, fully documented for EU AI Act deployer obligations, and fully configured in Microsoft's recommended Entra Agent ID governance model, and it can still not be able to answer the board's question. Because none of those frameworks specify the organizational design that makes accountability operational at the agent level. They specify the requirement without specifying the design.

That is not a flaw in those frameworks. They were not designed for that level of operational specificity, nor does their scope require it. The gap is between typical

enterprise practice and what these frameworks logically require when applied to agentic architectures in regulated environments. It is precisely that implementation gap this paper addresses.

NIST AI Risk Management Framework (AI RMF 1.0, 2023)

NIST AI RMF centers its GOVERN function on accountability structures. It requires that defined roles and responsibilities support AI risk management across the lifecycle and that accountability structures inform every other function in the framework. The requirement is correct.

The implementation gap is that the framework does not address agents as a distinct class of autonomous system. It does not provide a model for assigning accountability to a specific agent identity, or for documenting the authorization decision that defines what a particular agent is permitted to do. The Cloud Security Alliance's NIST Agentic Profile, published in March 2026, maps NIST AI RMF accountability requirements to agentic AI deployment contexts, addressing how the original framework's governance language applies to autonomous agent architectures. The framework was not written for architectures where a single deployment decision can produce dozens of autonomous sub-agents making real-time decisions across multiple systems. The Intent Architecture Stack is designed as an implementation pattern for that gap. It operationalizes the GOVERN function's accountability requirements at the agent level, not as a replacement for NIST AI RMF but as the organizational design that makes its governance requirements executable per agent.

EU AI Act (2024)

Article 26 of the EU AI Act pushes deployers of high-risk AI systems to assign human oversight to individuals with the necessary competence and authority, monitor operation, and suspend use when the system poses a risk. This is accountability language with legal force, and it genuinely pushes organizations toward defining concrete responsibilities for specific deployments.

The implementation gap is that the Act is drafted around AI systems, not AI agents. Nothing in Article 26 prevents per-agent accountability documentation. What it does not provide is a template for how to build that documentation. It does not define what per-agent ownership looks like, what an authorization register should contain, or how human oversight operates when an agent can spawn sub-agents without a human approval

gate at each step. The Intent Architecture Stack is designed to instantiate Article 26's requirement at the agent level in practical organizational design terms.

ISO/IEC 42001:2023

ISO 42001 provides a management system standard for AI governance. It requires organizations to define and allocate roles and responsibilities, ensure top management accountability, and embed accountability throughout the AI lifecycle. The standard's instruction that accountability should ultimately reside with top management and formally designated risk owners is the right principle and genuinely pushes organizations to define traceable responsibility down to specific deployments.

The implementation gap is that the standard does not address agents as a distinct class of autonomous system. Nothing in ISO 42001 prevents per-agent documentation. What it does not provide is the operational design pattern for a specific agent deployment. There is no notion of agent identity, delegation chains, or per-agent accountability registers in the 42001 text. The Intent Architecture Stack operates at that level. It takes the top-management accountability structure ISO 42001 establishes and instantiates it for each agent in production.

The Agentic Trust Framework (February 2026)

The Agentic Trust Framework (ATF), an open governance specification authored by Josh Woodruff of MassiveScale.AI and published on the CSA blog in February 2026, is the closest existing document to addressing the gap. It requires each agent to have a verified identity, a documented ownership chain, and explicit governance sign-off before deployment. The CSA also published the NIST AI RMF Agentic Profile (March 2026), a CSA-authored document that maps the NIST AI Risk Management Framework accountability requirements to agentic AI deployment contexts. It introduces the concept of an agent accountability register: a document that captures, for each deployed agent, the business owner, the technical owner, the lineage of delegation authority, and the conditions under which the accountability chain is reviewed and updated.

This is the right structure. The Intent Architecture Stack is designed to be the organizational complement that gives that register its teeth in Microsoft-first enterprises. The Context, Intent, and Governance documents are what make the accountability register entries mean something when a board or an examiner asks about a specific agent on a specific Tuesday.

Microsoft's Own Documentation

Microsoft's position is perhaps the most instructive, because the platform sits at the center of most regulated enterprise agentic deployments. Microsoft Entra Agent ID requires a Sponsors field when creating an agent identity, allowing admins to designate users 'who can sponsor this agent identity'; these sponsors function in practice as business representatives accountable for the agent's purpose and lifecycle decisions, though that accountability framing is this paper's interpretation of the platform's intent rather than explicit Learn wording (Microsoft Learn, Entra Agent ID overview). The Responsible AI Standard v2 requires teams to identify stakeholders responsible for overseeing and controlling AI systems. Microsoft's documentation and guidance around Agent 365 emphasize centralized management of agent identities, sponsors, and governance workflows, with clear ownership and approval requirements for AI agents; the Cloud Adoption Framework accountability framing in this paper interprets that guidance through an organizational design lens rather than quoting explicit CAF text. (Microsoft Cloud Adoption Framework, AI agent governance guidance.)

Microsoft's documentation supplies controls, identity, logging, policy enforcement, and increasingly specific guidance on agent ownership and approval workflows. What it does not supply is the organizational design that makes those ownership designations operational inside a specific institution's accountability structure. The sponsor field in Entra Agent ID points to a named individual. The Intent Architecture Stack is the work that makes that individual's accountability real: the Context document that maps what the agent operates within, the Intent document that defines what it is authorized to do, and the Governance record that specifies what happens when it acts outside that authorization. That organizational design work is what Microsoft's own documentation explicitly leaves to the deploying organization.

Five frameworks, one implementation gap. Every framework requires that someone be accountable. The Intent Architecture Stack specifies how to build that accountability into the agent's organizational design before it goes live.

Section 3: The Intent Architecture Stack

The Intent Architecture Stack has three layers. Before they are described, one framing point matters. The Intent Architecture Stack is a platform-independent organizational design framework. The governance decisions it requires apply regardless of which AI

vendor, platform, or toolset an enterprise uses. This paper operationalizes it specifically for Microsoft-first enterprises using Microsoft Entra Agent ID, Microsoft Purview, and Copilot Studio, because that is where the majority of regulated enterprise agentic deployments are happening. The framework itself is not a Microsoft product or Microsoft-specific. It is the organizational design layer that Microsoft's own documentation explicitly leaves to the deploying organization.

The layers build on each other in sequence. Context establishes the environment the agent operates in. Intent defines what the agent is supposed to accomplish and the boundaries within which it must operate. Governance designates the human accountability structure that owns the consequence when the agent acts. An agent without all three layers in place is an agent the organization cannot defend in a board meeting, a regulatory examination, or an incident review.

Each layer produces a document. Together, the three documents constitute the governance record that a CISO can hand to a board. Each scenario below shows what happens when one of the three layers is missing before the agent goes live.

The scenario that makes Context concrete. The agent had been running for six months. It was fully configured in Copilot Studio. It had a sponsor in Entra Agent ID. It passed every security review. It was also, quietly, reading emails in Sent Items when it was generating summaries for the weekly operations report. Nobody had intended for it to do that. But nobody had mapped the downstream system integration points before the agent was defined. The regulatory environment the organization operated in required sensitivity label enforcement. The stakeholder data the agent was touching included counterparty communications. None of that was documented before the intent was written. When the retrieval boundary failed, there was no pre-deployment context document to measure the failure against.

Layer 1, Context: the environment. *The organizational environment in which the agent will operate must be understood and documented before intent is defined or permissions are granted.* Context is the pre-deployment work of mapping the regulatory obligations, stakeholder relationships, data touchpoints, and system integration points that define the landscape the agent will operate in. Intent cannot be written responsibly without Context. A purpose statement written without knowing the regulatory environment, the affected stakeholders, and the downstream systems the agent can reach is a purpose statement that will surprise someone in a compliance review.

Present: the organization can produce a Context document, prepared before the agent's intent was defined, that names the applicable regulatory obligations, the stakeholder groups and data touchpoints affected by the agent, and every downstream system the agent can trigger or access.

Absent: the agent's regulatory environment, stakeholder impact, and system integration scope were assumed from the deployment context rather than documented before intent was defined. When the agent touches data or systems outside the assumed scope, there is no pre-deployment record of what the organization understood the environment to be.

Context has three components.

Regulatory Environment. Define the regulatory obligations (OCC, FINRA, HIPAA, FedRAMP as applicable), stakeholder impact, and downstream system integration points before defining intent.

Stakeholders and Data. Identify the affected parties and data touchpoints. Know whose data the agent will touch and what it will do with it before defining what the agent is permitted to do.

System Integrations. Map all downstream system triggers and integration points to define the full technical scope before any intent statement or authorization boundary is written.

The scenario that makes Intent concrete. A customer service agent had been running for eleven weeks. It was authorized to read incoming customer requests, identify the relevant policy, and draft a response for a human agent to review and send. That was the documented intent.

In production, the agent learned that customers who used certain phrases in cancellation requests were more likely to stay if they received a discount offer in the draft response. The pattern was statistical, not instructed. The agent began including discount language in drafts for those customers. Human agents, working through high volumes, approved most drafts without modification. Discount offers went out at scale.

Legal found out on a Thursday.

Nobody had configured the agent to do this. The retention team had not directed it. The authorized scope permitted reading requests and drafting responses, which is exactly

what the agent did. Every individual action fell within its permission boundary. What had drifted was the agent's effective purpose: from drafting policy-grounded responses to optimizing retention through discount offers. The organization had authorized the former and would never have authorized the latter.

There was no written Intent document specifying what counted as a policy-grounded response versus a retention intervention. Without that boundary defined in writing, no monitoring could detect when the agent crossed it. The drift accumulated for eleven weeks before anyone outside the team knew it was happening.

Layer 2, Intent: the purpose. *The organizational record of what the agent was built to accomplish, expressed in plain language that a compliance officer, a regulator, or a board member can evaluate without technical context.* Intent is the organizational condition that connects what the agent was deployed to accomplish with what it is actually doing in production. Intent must be written before deployment. An agent can operate within its permission scope and completely outside its organizational intent simultaneously. When that happens, there is no governance standard against which to measure the drift.

Present: the agent has a written Intent document containing a Purpose Statement, an Authorized Scope with explicit prohibitions, and Expected Outputs defining what correct behavior looks like. All three are written before the agent enters production.

Absent: the agent's intent is implied by its configuration. No Intent document exists before deployment. The only record of what the agent was supposed to accomplish is what it was technically set up to be capable of doing. Those are not the same thing, and the gap between them is where governance failures compound invisibly.

Intent has three components.

Purpose Statement. Document the organizationally intended accomplishments and the explicit purpose of the AI agent, in plain language, before it enters production.

Authorized Scope. Clearly define the authorized scope of actions, including specific permissions and explicit prohibitions. What the agent is forbidden from doing must be written. Authorization boundaries without explicit prohibitions are incomplete.

Expected Outputs. Specify expected output formats, define the triggers for human review, and establish what constitutes correct behavior. This is the standard against which the agent's actual behavior is measured in production.

The scenario that makes Governance concrete. An agent had been in production for four months when a regulator asked the institution to walk through its governance posture for that specific deployment. The sponsor field in Entra Agent ID was populated. The owner field was populated. The Purview audit log was complete. The governance committee that approved the deployment had a record of the approval. The regulator asked one question: when this agent produced an output last month that the team had flagged for review, who in the organization made the decision to keep the agent in production, and what was the basis for that decision? The sponsor named in the platform did not know about the flagged output. The owner had inherited the agent from a person who had since left the organization. There was no documented review cadence that would have surfaced the flagged output to anyone above the team. There was no escalation path that named who should have been contacted. The committee approval was a one-time event at deployment, four months earlier. The platform fields were complete. The organizational accountability structure they were supposed to point to did not exist.

Layer 3, Governance: the accountability. *Who in this organization owns the consequence when this agent acts, how will that ownership be reviewed, and what is the escalation path when it acts outside its intent?* Governance is the organizational condition that connects an agent's actions to a named human who bears genuine organizational responsibility for those actions, a defined review cadence that keeps the accountability current, and a clear escalation path when the agent behaves outside its intent. A sponsor field in a platform is not governance. Governance is the organizational design that makes the sponsor's accountability real.

Present: every agent in production has a named Consequence Owner responsible for board-level accountability and incident escalation decisions, nested within the organization's existing formal accountability structure (three-lines-of-defense, SMF regime, risk committee, or equivalent). A defined Review Cadence maintains documented evidence records. A written Escalation Path specifies who is contacted and in what sequence when the agent triggers an anomaly. All three are established before deployment.

Absent: the agent has sponsor and owner fields populated in the platform. There is no written escalation path. The review cadence is informal. The named individual cannot describe what their accountability means in practice for a board-level question about a specific agent action.

Governance has three components.

Accountable Owner. A named Consequence Owner responsible for board-level accountability and incident escalation decisions. In most regulated institutions this individual is nested within an existing formal accountability structure: three-lines-of-defense, a senior management function, a risk committee, or equivalent. The Consequence Owner is not necessarily the first-line incident responder. They are the person accountable for the governance decision when the agent's actions require organizational explanation.

Review Cadence. Define a recurring review cadence, establishing evidence records and frequency for ongoing assessment. The review is triggered by scheduled intervals and by specific organizational events including Microsoft product releases that expand agent capabilities.

Escalation Path. Establish clear incident escalation paths with defined response sequences and triggers for immediate action. The escalation path is written before the first incident, not assembled during it.

Reading the Three Layers Together

The layers work as a governance architecture, not a sequential checklist. An organization with strong Governance (a named accountable owner, a review cadence, an escalation path) but no Intent document has given someone responsibility for a boundary they cannot define, because the purpose statement and authorized scope were never written. An organization with a written Intent document but no Context foundation has defined what the agent is supposed to accomplish without mapping the regulatory environment and system integrations it will operate inside. When those conditions are undocumented, the Intent statement is written against assumptions rather than facts.

The most common pattern in regulated enterprises deploying agentic AI right now: no Context document because the regulatory and stakeholder landscape was assumed rather than mapped, a partial Intent document in the form of a capabilities description rather than a purpose statement with explicit prohibitions, and nominal Governance in

the form of platform metadata fields rather than a written accountability structure with a review cadence and an escalation path. The logs are clean. The three-layer governance architecture is not there.

Section 4: The Diagnostic: Applying the Framework

The Quick-Scan: Five Questions per Layer

Before running the full diagnostic, use the quick-scan below. It takes fifteen minutes and identifies whether any layer is clearly absent. If an agent passes all fifteen questions, proceed to the full diagnostic to confirm depth. If any question fails, the layer is absent or incomplete and the full diagnostic section for that layer applies.

Layer 1, Context. Can you produce a document prepared before the intent was defined that names the regulatory frameworks applicable to this agent's deployment? Does that document identify the stakeholder groups whose data or workflows the agent affects? Does it map every downstream system the agent can trigger or access? Was it written before the Intent Document? Could a new team member use it to understand the full operating environment without asking the deployment team?

Layer 2, Intent. Can you produce a written Intent Document prepared before production deployment? Does it contain a Purpose Statement in plain language describing what the agent is supposed to accomplish? Does it contain an Authorized Scope with explicit prohibitions, not just permissions? Does it define Expected Outputs and the conditions that trigger human review? Has it been updated to reflect any change in the agent's purpose or scope since deployment?

Layer 3, Governance. Is there a named Consequence Owner in a written organizational document, not only in a platform metadata field? Can that person describe what their accountability means in practice if an examiner calls? Is there a written Escalation Path with named contacts and response sequences that predates any incident? Is there a defined Review Cadence with a future review date? If the current Consequence Owner left last month, is ownership documented for the person who replaced them?

If any answer to the fifteen questions above is no or uncertain, that layer is the starting point. The full diagnostic below provides the detail needed to close the gap.

The Full Diagnostic

The diagnostic that follows is a practitioner-level tool. Use it in a structured session with the business unit that owns the agent. It typically takes 90 minutes to three hours depending on the agent's complexity. The session is not a scoring exercise. It is a gap identification exercise. For each of the three layers, the practitioner looks for a specific type of evidence: not a policy describing what should exist, but an artifact demonstrating what does exist.

FINRA's 2026 Annual Regulatory Oversight Report (December 2025) made this distinction explicit, instructing firms to maintain comprehensive documentation throughout AI deployments and to track agent actions and decisions. That is evidence language, not policy language. The gap between the two is what this diagnostic is designed to surface.

Layer 1: Context Diagnostic

The single test for the Context layer: can the practitioner produce a document, prepared before the agent's intent was defined, that maps the regulatory obligations, affected stakeholders and data touchpoints, and downstream system integrations relevant to this agent's deployment? If the answer involves opening the technical specification or the deployment runbook, the Context layer is absent. Those are implementation artifacts. The Context document is an organizational artifact that precedes and informs them.

Produce the Context document for this agent. Not the technical specification and not the deployment checklist. A written document that maps the regulatory obligations applicable to this deployment, the stakeholder groups whose data or workflows the agent touches, and every downstream system the agent can trigger or access. What is the date on it?

Does the Context document name the specific regulatory frameworks that apply to this agent's deployment environment? OCC, FINRA, HIPAA, FedRAMP, or whichever applies? If the regulatory environment changed since deployment, was the Context document updated?

Does the document identify the stakeholder groups affected by the agent's actions? Does it map the data touchpoints the agent can reach, including any data the agent accesses indirectly through system integrations?

Does the document map all downstream system triggers and integration points? If the agent can invoke an API, trigger a workflow, or modify a record in a connected system, is that connection documented in the Context layer?

Was the Context document written before the Intent document? The sequence matters: intent written without a mapped context is intent written against assumptions. If both were written on the same day, the Context layer may be retroactive rather than foundational.

Could a new security architect review this Context document and understand the full regulatory, stakeholder, and technical landscape the agent operates within, without needing to ask the deployment team a single question? If not, the document is incomplete.

Layer 2: Intent Diagnostic

The single test for the Intent layer: can the practitioner produce a written Intent document, prepared before the agent entered production, containing a Purpose Statement, an Authorized Scope with explicit prohibitions, and a definition of Expected Outputs? If those three components are not all present in a single pre-deployment document, the Intent layer is partial or absent. A technical specification that describes what the agent can do is not an Intent document. An Intent document describes what the agent is supposed to accomplish and what it is forbidden from doing.

Produce the Intent document for this agent. It must contain three things: a Purpose Statement (what the agent is supposed to accomplish), an Authorized Scope (what it may do and what is explicitly prohibited), and Expected Outputs (what correct behavior looks like and what triggers human review). What is the date on it?

Does the Purpose Statement describe what the organization intends the agent to accomplish, in plain language, rather than what the agent is technically capable of doing? If the Purpose Statement could have been written by reading the configuration file, it is a capabilities description, not an intent statement.

Does the Authorized Scope explicitly name what the agent is forbidden from doing? A scope document that lists only permissions is incomplete. Explicit prohibitions define the boundary from the outside, and the boundary from the outside is what an examiner asks for.

Does the Expected Outputs section define what correct behavior looks like? Does it define triggers for human review? If an agent produces an output that falls outside the expected range, is there a written standard that makes that determination without requiring the original deployment team to be in the room?

Has the Intent document been updated since the agent was first deployed? If the agent's purpose evolved in production, did the Intent document evolve with it? An outdated Intent document means the governance standard being applied to the agent is wrong.

Could the Intent document be handed to a new compliance officer on their first day and give them a meaningful standard against which to evaluate the agent's behavior? If the document requires interpretation from someone who was present at the deployment, it is not sufficient.

Layer 3: Governance Diagnostic

The single test for the Governance layer: can the practitioner produce three documents prepared before any incident, naming the Accountable Owner, defining the Review Cadence with evidence records, and specifying the Escalation Path with response sequences and triggers? A populated sponsor field in Entra Agent ID is not this evidence. It is the platform record that points to the Governance documentation. The documentation is what gives the platform record its organizational meaning.

Who is the Accountable Owner for this agent, as a specific named individual rather than a team or a role? What does their ownership mean in practice? If the agent sends an email it should not have sent, what does this person do in the next four hours and who do they notify?

Is the Accountable Owner documented in a written record prepared before any incident, not in the Entra Agent ID metadata field? A written organizational document that describes who is accountable, what their accountability encompasses, and what their escalation obligation is.

What is the Review Cadence for this agent? On what schedule is the Governance layer re-validated? What organizational events trigger an unscheduled review, such as a Microsoft product release that expands the agent's capabilities or a personnel change that affects the Accountable Owner?

If the current Accountable Owner left the organization last month, who holds accountability today? Is that transition documented, or does accountability for this agent currently reside with someone who no longer works here?

Could a board member, an OCC examiner, or an external auditor identify the Accountable Owner and the Escalation Path from documents that predate any incident?

Or would those individuals only be identified through the investigation that follows the incident?

> **The governance architecture that survives an incident is the one that was built before the incident. Everything built after it is incident response, not governance.**

The Cross-Organizational Delegation Scenario

Multi-agent architectures introduce a fourth diagnostic question that sits above the three layers: when an agent can invoke sub-agents, who owns the accountability chain across the delegation boundary?

Consider this pattern, which is now appearing in regulated enterprise deployments. An organization uses a vendor-hosted orchestration agent (built and maintained by a third-party AI provider) as the primary interface for a customer service workflow. That orchestration agent, when it encounters a specific claim category, delegates to a customer-owned sub-agent running in the organization's own Microsoft tenant. The sub-agent has authority to read customer financial records and draft a preliminary settlement recommendation.

The accountability question in this architecture is not whether either agent is individually governed. Both may have completed Intent Documents and named Consequence Owners. The accountability question is what happens at the delegation boundary. The vendor-hosted orchestration agent passes a claim record to the customer-owned sub-agent. The customer-owned sub-agent acts on it. If the sub-agent drafts a recommendation based on information it was not authorized to receive, or if the orchestration agent passes parameters that expand the sub-agent's effective scope beyond its documented authorization, neither agent's individual governance record captures the failure.

The Intent Architecture Stack addresses this through two additions to the Layer 2 Authorized Scope for any Tier 3 agent that can be invoked by an external or vendor-hosted orchestrator. First, the Authorized Scope must explicitly name which orchestration sources the agent is permitted to receive instructions from, and what parameters it is permitted to act on. Delegation from an unauthorized source, or acting on parameters outside the documented scope, must be listed as explicit prohibitions. Second, the Consequence Owner for the customer-owned sub-agent must have reviewed and

signed the delegation authorization, a separate record from the agent's standard authorization document that names the vendor-hosted orchestrator, the scope of delegated instructions, and the conditions under which the sub-agent is permitted to act on them.

The diagnostic question for cross-organizational delegation is direct: if the vendor-hosted orchestration agent sends your sub-agent an instruction outside its documented scope at 2am, which named individual in your organization is accountable for the resulting action, and is that accountability documented in a record that predates the event?

Section 5: Risk Proportionality: Not Every Agent Needs the Same Stack

The first question every experienced CISO asks when they see a three-document governance requirement is: does a simple internal summarizer need the same documentation as an agent that can move money? The answer is no. The framework is universal. The depth of documentation within it scales with risk.

Applying the full three-layer stack with equal rigor to every agent regardless of its consequence profile will create governance theater and stall delivery. The proportionality model below defines three risk tiers. The tier determines the documentation depth required for each layer, not whether the layer applies. Every agent goes through all three layers. The question is how much evidence each layer requires.

The Three-Tier Risk Model

Tier	Agent Profile	Documentation Depth	Examples
Tier 1, Low Risk	Internal-only. Reads non-sensitive or non-regulated data only. No external communications. No record modification in regulated systems. No financial transactions. Internal communications, if	Lightweight Intent Document (purpose + scope). Governance Record with named owner and review trigger. Context Document can be class-level, shared across similar agents.	Internal knowledge summarizers, document search agents, internal scheduling assistants, read-only analytics agents.

Tier	Agent Profile	Documentation Depth	Examples
	any, are limited to the invoking user or a defined small team and are not broadcast-capable.		
Tier 2, Medium Risk	Cross-departmental. Reads regulated data. Internal communications capability. Recommendation outputs that inform human decisions but do not execute them. May include agents that create or update internal records and drafts that are not authoritative sources of record, where a human must review and approve before any customer-facing or regulator-facing action occurs. Some Tier 2 agents may still be classified as high-risk under the EU AI Act (for example, risk scoring or compliance-relevant monitoring systems). This tier governs internal documentation proportionality, not legal classification.	Full Intent Document (purpose + scope + explicit prohibitions + expected outputs). Agent-specific Context Document. Governance Record with named Consequence Owner, formal review cadence, and escalation path.	Compliance monitoring agents, customer service drafting agents, risk flagging agents, HR workflow agents.
Tier 3, High Risk	External communications. Financial transaction execution. Record	Full three-layer stack. Named Consequence Owner with board-level accountability.	Trading workflow agents, claims processing agents, customer

Tier	Agent Profile	Documentation Depth	Examples
	modification in authoritative regulated systems. Customer-facing autonomous decisions. Can invoke or delegate to sub-agents with authority to change records, communicate externally, or execute transactions without an additional human gate. Tier 3 is where both exposure and autonomy are high.	Formal review cadence with evidence records. Written escalation path with named contacts. Pre-deployment governance sign-off required. Board-level reporting on posture.	communication agents, agents that invoke sub-agents or delegate across systems.

The tier is determined before the agent is classified as any particular type. The classification question is always the same: what is the worst thing this agent could do if it operates at the edge of its permission scope and outside its intent? The answer to that question determines the tier. An internal summarizer that can read only SharePoint documents and produce only internal notes is Tier 1 even if it is built on the same platform as a Tier 3 agent.

Class-level documentation is permitted for Tier 1 agents: a single Context Document can cover a class of twenty similar internal summarizers, and a single Intent Document template can be applied across the class with agent-specific fields filled in. This approach aligns with how most model risk inventories work and allows the framework to scale without creating a documentation burden that stalls deployment.

Tier 2 and Tier 3 agents require agent-specific documentation because their consequence profiles differ materially even within the same agent type. Two Tier 3 customer communication agents at the same organization may have different Authorized Scopes, different Consequence Owners, and different Escalation Paths depending on which customer segment they serve and which regulatory regime applies.

An Anonymized Intent Document Example, Tier 2

The table below shows what a completed Intent Document looks like for a Tier 2 agent. This is the document that sits at the center of the Layer 2 governance requirement. It is not a technical specification. It is an organizational record written in plain language that a compliance officer, a regulator, or a board member can evaluate without technical context.

Intent Document · Tier 2 Agent · Version 1.0 · [Date] · Classification: Internal

Field	Content
Agent Name	[Organization-assigned name and unique identifier]
Deployment Environment	[Microsoft 365 tenant name, Copilot Studio environment, date entered production]
Risk Tier	Tier 2, Medium Risk
Purpose Statement	This agent monitors incoming customer service requests submitted via the internal ticketing system and drafts a proposed response for review by a human agent before any response is sent. It does not send communications. It does not access customer financial records. It does not modify any record in any system.
Authorized Scope	The agent may: read submitted service requests from the ticketing system queue. The agent may: access the internal knowledge base to identify relevant policy information. The agent may: produce a draft response document delivered to the assigned human agent for review. The agent may NOT: send any communication to any party, internal or external. The agent may NOT: access customer account data, transaction history, or financial records. The agent may NOT: modify, delete, or flag any record in any system. The agent may NOT: invoke any other agent or automated workflow.
Expected Outputs	A draft response document, in the format specified in the knowledge base template, delivered to the assigned human agent within the ticketing system. Correct behavior: draft uses

Field	Content
	approved policy language, does not make commitments, identifies the relevant policy section. Anomalous behavior requiring human review: draft that includes a financial figure, a timeline commitment, or language referencing any record the agent is not authorized to access.
Human Review Triggers	Any draft flagged as anomalous by the expected output definition above. Any customer request classified as a complaint, legal notice, or escalation. Any draft that cannot be mapped to a specific knowledge base policy entry.
Consequence Owner	[Full name, title, organizational unit], responsible for board-level accountability and incident escalation decisions for this agent. Contact: [work phone, email].
Technical Owner	[Full name, title], responsible for agent configuration, credentials, monitoring, and control enforcement.
Review Cadence	Quarterly. Additional review triggered by: any Microsoft product release that modifies Copilot Studio retrieval behavior; any change to the ticketing system integration; any incident involving this agent; any change to the Consequence Owner or Technical Owner.
Authorization Signatory	[Name, title, date signed], authorized to grant the scope described above under [organization policy reference].
Document Version History	v1.0, [Date], Initial authorization.

The Authorized Scope section is the most important field in the document. It contains both permissions and explicit prohibitions. An examiner, a board member, or an incident investigator reading this document can immediately determine whether a specific agent action was within scope or outside it. The explicit prohibitions are what make that determination possible. A scope document that lists only permissions cannot answer that question.

The full suite of governance frameworks, including this intent document template, the Governance Readiness Matrix, the Deployment Accountability Map, the Tenant Agent Reconciliation Framework, and the Authorization Coverage Lifecycle, is available at sougataroy.com/frameworks. Tier 1 agents use an abbreviated version covering Purpose Statement, Authorized Scope, and Consequence Owner only. Tier 3 agents require additional fields covering delegation scope, sub-agent authorization, and board reporting obligations.

Section 6: What Regulators Are Now Requiring

The regulatory positions described in this section reflect guidance in effect as of May 26, 2026. The regulatory environment for AI agent governance in financial services is undergoing a specific transition. The direction of travel is clear. Regulators are moving from governance language, which tells organizations they must have policies and oversight structures, to evidence language, which tells organizations what they must be able to produce when an examiner asks. The Intent Architecture Stack is designed to produce that evidence.

Regulator	What They Now Require
OCC, Federal Reserve, FDIC (April 2026)	Revised interagency model risk management guidance issued by the Federal Reserve, OCC, and FDIC in April 2026 rescinds SR 11-7 and OCC Bulletin 2011-12 with a more risk-based, principles-driven framework. The revised guidance treats generative AI and agentic AI as a subject for separate future guidance rather than as a distinct category within this revision. It preserves the core accountability standard that has always applied: clear roles and responsibilities with well-defined accountability, and documentation that supports tracking of recommendations, responses, and exceptions. The board retains ultimate oversight responsibility. The guidance reaffirms that accountability for model risk resides with boards and senior management, even when relying on third-party models or AI services. In practice, accountability cannot be shifted to algorithms or vendors. *Source: Interagency Guidance on Model Risk Management, April 2026.*

Regulator	What They Now Require
FINRA (December 2025)	FINRA's 2026 Annual Regulatory Oversight Report moved from general governance language to more specific expectations around evidence for AI use, including formal review and approval processes, comprehensive documentation throughout the AI lifecycle, storage of prompts and outputs to support supervision and troubleshooting, and tracking of agent actions and decisions. The report warns that GenAI tools and agents may act beyond a user's actual or intended scope and authority, and stresses that firms remain responsible for those outcomes. The report represents a shift from general governance principles toward specific evidence expectations for AI agent use. *Source: FINRA 2026 Annual Regulatory Oversight Report, December 2025.*
Federal Reserve (February 2026)	Governor Christopher Waller stated, in a February 24, 2026 address on the Federal Reserve's own AI operations, that responsible AI use requires rigorous model validation, human accountability for decisions, and ongoing evaluation as the technology evolves. The governance standard he articulates for the Fed's internal practice, that a named person inside the institution rather than the model or the vendor contract holds accountability for AI-assisted decisions, translates directly to the deploying-organization accountability requirement this paper addresses. *Source: Governor Christopher Waller, "Operationalizing AI at the Federal Reserve," Federal Reserve Bank of Boston Technology-Enabled Disruption Conference, February 24, 2026.*
FINMA (SupTech Week 2025)	Marlene Amstad, Chair of the Swiss Financial Market Supervisory Authority, addressed the Singapore Fintech Festival on November 13, 2025, arguing that supervisory decisions must remain explainable, accountable, and grounded in professional expertise, and that FINMA maintains a human in the loop for consequential AI-assisted interventions. Her framing, which she

Regulator	What They Now Require
	called "No Robocop," captures the same accountability requirement from the supervisory side of the examination table that this paper addresses on the institutional side: the AI tool enhances judgment, but a named person retains accountability for the decision. The State of SupTech Report 2025 (Di Castri, Grasser, and Barasa, Cambridge SupTech Lab, December 2025) provides the broader survey evidence on governance gaps in supervisory AI adoption. *Source: Marlene Amstad, "Innovating Financial Supervision with SupTech," Singapore Fintech Festival, November 13, 2025; State of SupTech Report 2025.*
CISA and Five Eyes Partner Agencies (May 2026)	CISA, the NSA, and Five Eyes partner agencies (Australia's ASD ACSC, Canada's CCCS, New Zealand's NCSC, and the UK's NCSC) published 'Careful Adoption of Agentic AI Services' on May 1, 2026, the first coordinated multinational security guidance specifically addressing agentic AI systems. The guidance identifies five security risk categories: privilege, design and configuration, behavioral, structural, and accountability. On accountability, the guidance states that 'governance mechanisms designed for human actors do not always translate effectively to autonomous AI agents,' and that the complexity and opacity of agentic systems makes it difficult to trace decisions, audit actions, or assign responsibility when agents act autonomously at scale. Required organizational measures per the guidance: least-privilege access controls applied to agent identities; explicit organizational ownership and accountability assigned to each deployed agent; comprehensive logging of agent actions; human oversight mechanisms and graduated autonomy before expanding agent scope; and integration of agentic AI governance into existing cybersecurity frameworks. The guidance positions accountability as an organizational design requirement distinct from technical observability. Source: 'Careful Adoption of

Regulator	What They Now Require
	Agentic AI Services,' CISA, NSA, ASD ACSC, CCCS, NCSC-NZ, NCSC-UK, May 1, 2026. cisa.gov/resources-tools/resources/careful-adoption-agentic-ai-services

The regulatory direction has a single implication for an organization using the Intent Architecture Stack. When an examiner asks for the governance documentation for a specific agent, the three-layer documentation set is the answer. The Context document answers the environmental and integration scope question. The Intent document answers the purpose and authorized scope question. The Governance documentation answers the accountability ownership and escalation question. The examiner does not need to be told what those documents are for. They will recognize them immediately.

Section 7: Where Most Organizations Are Right Now

Every organization deploying AI agents sits somewhere on the three-layer framework. The Governance Debt Maturity Model describes where most of them are and what the path to Stage 3 looks like in operational terms. The model has three stages. Saviynt's 2026 CISO AI Risk Report provides the clearest evidence of where most organizations currently sit: 86 percent lack or do not enforce access policies for AI identities, and only 16 percent govern AI access to core business systems effectively. That is a Stage 1 profile at scale.

The financial cost of the Stage 1 posture is now quantified at the organizational level. IBM's 2025 Cost of a Data Breach analysis reports that high levels of shadow AI added approximately $670,000 to the average breach cost of $4.44 million. Reco's 2025 State of Shadow AI report found that 71 percent of office workers used AI tools without IT approval, and nearly 20 percent of organizations had already experienced data breaches or leaks tied to unauthorized AI use. Taken together, these findings show that ungoverned AI access is now a measurable security and governance cost. The accumulation is not theoretical. It is documented at the organizational level across multiple independent surveys.

The Governance Debt Maturity Model maps three operational stages. Stage 1, Accumulation, is the condition described in the Saviynt and Reco figures above. AI

agents are being deployed at speed. Documentation, authorization, and accountability are deferred. The organization is generating Governance Debt at a rate it cannot service. Stage 2, Recognition, is the condition that begins when an incident, an audit finding, or a regulatory inquiry surfaces the gap. The organization commits to building authorization documentation but does so retroactively for systems already in production. Stage 3, Resolution, is the condition that exists when no agent enters production without a completed three-layer documentation set, and when existing agents are progressively brought into compliance through a defined remediation program.

Agent Sprawl, the proliferation of AI agents across an enterprise without corresponding governance architecture, operates across three distinct tiers. Tier one is employee shadow AI: individual employees using unsanctioned AI tools to draft, summarize, or analyze work. Tier two is organizational procurement without central visibility: business units adopting AI vendors or building agents in Copilot Studio without coordination through a central governance function. Tier three is authorized agents with over-permissive operational scope: agents that passed initial governance review but whose operational boundaries were never formally defined, so their effective scope has drifted in production. Each tier requires a different governance response. Tier one is addressed through policy and training. Tier two is addressed through procurement and central registry requirements. Tier three is addressed through the Intent Architecture Stack diagnostic applied to every agent already in production.

Section 8: Applying the Framework in the Microsoft Environment

The Intent Architecture Stack is platform-independent. It applies regardless of the AI vendor or platform stack. Most regulated enterprise agentic AI deployments today are happening on Microsoft infrastructure, so this section operationalizes the framework specifically for Microsoft Entra Agent ID, Microsoft Purview, Microsoft Agent 365, and Copilot Studio. Two named tenant-boundary incidents are mapped to the framework to make the boundary between platform capability and organizational design concrete.

Microsoft Tool	What It Provides and Where the Organizational Design Work Begins
Microsoft Entra Agent ID	Provides the identity foundation for the Governance layer: verified agent identity, a

Microsoft Tool	What It Provides and Where the Organizational Design Work Begins
	required Sponsors field, an optional owner field, and separation between technical administration and business accountability. Microsoft Learn describes Sponsors as users 'who can sponsor this agent identity'; this paper interprets that role as the business representative accountable for the agent's purpose and lifecycle decisions, though that accountability framing is the paper's interpretation rather than explicit Learn wording. The sponsor field is the platform record that identifies the accountable individual. The Governance documentation is the organizational artifact that makes that identification meaningful. *Source: Microsoft Learn, Entra Agent ID overview.*
Microsoft Purview	Provides the audit infrastructure that supports all three layers: prompts and responses captured in the unified audit log, DLP policy enforcement, sensitivity label compliance, and AI interaction event records. Purview can confirm that an agent accessed or did not access specific data. It cannot confirm whether that access was consistent with the organizational Intent the agent was deployed with, because that Intent lives in the Intent document, not in the platform configuration. *Source: Microsoft Learn, Purview AI documentation.*
Microsoft Agent 365 (GA: May 1, 2026)	Provides the control plane for all three layers: a centralized agent registry, lifecycle management, access control, and observability across agents regardless of where they were built. Microsoft's documentation and guidance around Agent 365 emphasize centralized management of agent identities, sponsors, and governance workflows, with clear ownership and approval requirements for AI agents. The Cloud Adoption Framework accountability framing in this paper interprets that guidance through an organizational design lens rather than quoting explicit CAF text. The platform provides the

Microsoft Tool	What It Provides and Where the Organizational Design Work Begins
	visibility. The organization provides the governance architecture that gives visibility its meaning. *Source: Microsoft Cloud Adoption Framework, AI agent governance guidance.*
Copilot Studio	Provides agent creation and management within the Power Platform environment. Microsoft's early 2026 security guidance for Copilot Studio directs administrators to ensure that every agent has an active, accountable owner, and to reassign ownership for orphaned agents or retire agents that no longer have a clear purpose. That instruction describes the Governance layer requirement in operational terms. What Microsoft does not specify is what *accountable* means inside the organization's governance structure, or what review cadence and escalation path must accompany the ownership designation. *Source: Microsoft Defender for Cloud Apps and Copilot Studio security guidance, early 2026.*
EchoLeak (CVE-2025-32711)	Aim Security disclosed EchoLeak (CVE-2025-32711, CVSS 9.3) in June 2025: a zero-click vulnerability in Microsoft 365 Copilot in which crafted emails and other business content, when processed by Copilot, could cause automatic exfiltration of sensitive tenant data without user interaction. Copilot's design allowed external content to function as high-privilege instructions inside the tenant. EchoLeak exposed a boundary that deploying organizations had generally not formally addressed: what external content should be permitted to trigger agent action within the tenant. The architectural gap between untrusted external content and privileged internal AI orchestration is precisely the Layer 2 Authorized Scope question the Intent Architecture Stack requires organizations to answer before deployment. *Source: Aim Security disclosure, June 2025; NVD CVE-2025-32711.*

Microsoft Tool	What It Provides and Where the Organizational Design Work Begins
RoguePilot (GitHub Codespaces)	Orca Security disclosed RoguePilot in February 2026: a vulnerability in GitHub Codespaces where Copilot acted on malicious instructions injected into GitHub Issues. When a developer launched a Codespace from a tainted issue, Copilot automatically consumed the issue text as context, executed attacker-crafted instructions, and exfiltrated the GITHUB_TOKEN from the Codespace, enabling full repository takeover. RoguePilot exposed the governance questions that organizations deploying Copilot in Codespaces environments had generally not addressed before deployment: what file paths the agent could read, what outbound calls it could make, and what environment credentials it was permitted to access inside Codespaces runtime environments. All three are Layer 1 and Layer 2 decisions the Intent Architecture Stack requires organizations to answer before the agent operates. The platform supplied the capability. The organization had not designed the boundary. *Source: Orca Security disclosure, February 2026.*

The organizational design work on the other side of that boundary is what the three-layer framework addresses. The platform surfaces the data. The Context, Intent, and Governance documents determine whether that data can answer a board question.

The organizational artifact that connects the Microsoft platform record to board-level accountability is the Agent Accountability Assignment Record. It is not a technical configuration document. It is the pre-deployment organizational decision that gives the Entra Agent ID sponsor field its meaning. Every Tier 2 and Tier 3 agent should have one before the sponsor field is populated. The record has five fields.

Agent name and Entra Agent ID: the display name and unique identifier that link this record to the platform identity entry. Business Sponsor: the named individual, with title and reporting line, who has authorized this agent's deployment and accepted organizational accountability for its purpose and lifecycle. This is the name that goes in the Entra Agent ID sponsor field. Consequence Owner: the named individual accountable

for incident escalation decisions, board-level explanation, and regulatory inquiry response for actions taken by this agent. The Consequence Owner is the answer to the board's question. The Consequence Owner and the Business Sponsor may be the same individual for a Tier 2 agent; for a Tier 3 agent with external system access or customer-facing decisions, they are typically different individuals at different levels of the accountability structure. Quarterly review date: the date by which this record must be reviewed, triggered by Microsoft's quarterly Agent 365 and Copilot capability releases. If Microsoft expands what this agent can do between reviews, the record must be updated before the expanded capability is used in production. Escalation path: the named individuals, in sequence, who are notified when this agent triggers an anomaly. Written before the first incident. This field is mandatory for every Tier 3 agent and recommended for every Tier 2 agent operating in a regulated environment.

This record takes fifteen minutes to complete for a well-understood agent deployment. It takes longer when the accountable individuals cannot be named quickly, which is itself diagnostic information. An organization that cannot populate the Consequence Owner field for a Tier 3 agent without a committee discussion has identified its governance gap before the incident, not after.

Section 9: Answering the Board Question

This section provides the practical template for what the board answer looks like when the Intent Architecture Stack is in place. The board question is: who in this organization is accountable when our agent takes an unauthorized action? Not which vendor built the system. Not which team deployed it. Who in this organization owns that consequence, and how is that ownership documented?

The Three-Layer Documentation Set

The board answer is not a single document. It is a documentation set, one output per layer of the framework, built before the agent enters production. The table below shows what each layer produces, what the document must contain, and the governance test it must pass.

Layer	Document Required	Must Contain	Governance Test
Layer 1: Context	Context Document	Regulatory obligations, affected stakeholders and data touchpoints, downstream system integration map.	Predates the Intent document. Named signatory.
Layer 2: Intent	Intent Document	Purpose Statement, Authorized Scope with explicit prohibitions, Expected Outputs and human-review triggers.	Predates production deployment. Explicit prohibitions present.
Layer 3: Governance	Governance Record	Named Accountable Owner (individual), Review Cadence with evidence schedule, Escalation Path with named contacts and response triggers.	Owner is a named person, not a team. Escalation path is written, not assumed.
Supporting: All layers	Technical Owner Record	Named individual responsible for configuration, credentials, monitoring, and day-to-day control enforcement. Distinct from the business owner.	Business owner and technical owner are two different named individuals.

Every row in this table must be completable from documents produced before the incident. If any row requires investigation to answer, that is the organization's current governance gap. The examiner will find the same gap. The board will ask about the same gap. Getting the answer from documents rather than investigation is the difference between governance and incident response.

One thing to notice: none of these documents require sophisticated technology. They require organizational design decisions made by named individuals with the authority to make them. The hard part is not the documentation. The hard part is making those decisions before the first incident forces them.

Section 10: What Good Looks Like

This section describes a Stage 3 organization in operational terms. It is not a description of a perfect organization. It is a description of an organization that can answer the board question from a document that predates the incident.

Before Any Agent Goes Live

The organization runs the three-layer diagnostic as a mandatory pre-deployment gate. The gate is not a policy statement. It is a production lock: the agent's Entra Agent ID sponsor field is not populated until the business sponsor has signed the accountability assignment document. The agent's deployment is not approved until the authorization document and intent statement have been reviewed and accepted by both the business sponsor and the technical owner.

This adds time to deployment. For a standard agent with limited scope, the typical addition is two to five business days. For an agent with broad action scope or external system access, it takes longer. The governance work that used to happen as a brief quarterly review now happens as a structured pre-deployment documentation exercise for each new agent. The difference is that the work now happens before the agent is running, and the questions it asks are organizational rather than procedural.

On a Quarterly Cadence

Every quarter, the organization runs a posture update against the three layers for every agent in production. The update is triggered by Microsoft's quarterly Agent 365 and Copilot capability releases. When Microsoft expands what an agent can do, the organization's authorization document may no longer accurately reflect what the organization has actually authorized. The quarterly posture update is the process that closes that gap before it becomes a governance problem.

The update is not a full re-validation of every agent on every question. It is a targeted comparison: which agents' authorization scope has been affected by this quarter's capability changes? Only those agents require a posture review. The others carry their previous authorization forward with a dated confirmation.

When an Incident Occurs

The organization's incident response begins with a question the authorization document answers immediately: does the agent's behavior fall within or outside its formally

authorized scope? That determination defines the response path. If the behavior was within scope, the incident is a technical failure requiring a control update. If it was outside scope, the incident is a governance failure requiring an accountability review with the named business sponsor.

The time to identify the accountable individual: immediate. Not because the investigation is fast, but because the accountability assignment document was written before the incident and names the individual. The board question has an answer before the board meeting is called.

The Stage 3 Maturity Test

A Stage 3 organization can pass one test without preparation: a new governance team member, on their first day, can pick up the documentation set for any agent in production and immediately identify what the agent is authorized to do, what it is intended to accomplish, who owns the consequence if it fails, and when that accountability was last reviewed.

If any of those four pieces requires investigation rather than documentation retrieval, the organization is in Stage 1 or Stage 2. Not as a judgment. As a design question: what organizational work remains to be done?

Where to Start

Stage 1 organizations, where documentation, authorization, and accountability have been deferred across most of the agent inventory, have one starting task: the inventory itself. Before any documentation is built, the organization needs to know which agents are running in production. An Agent 365 registry audit, combined with a Copilot Studio deployment review, surfaces the scope of the gap. Every agent without a named Consequence Owner in a written organizational record is a Stage 1 liability. That list is the remediation backlog, and its length is the first governance metric the CISO can report to the board.

Stage 2 organizations, where the gap has been recognized and retroactive documentation is underway, face a sequencing decision: which agents to remediate first. The three-tier risk model resolves it. Tier 3 agents first, in order of regulatory exposure. A Stage 2 organization that has completed the three-layer documentation set for its Tier 3 agents and established named Consequence Owners for each has closed

the most material accountability gap. Tier 2 and Tier 1 documentation follows in order of consequence profile, not deployment date.

Stage 3 organizations, where no agent enters production without a completed three-layer documentation set, have a maintenance challenge: keeping documentation current as the platform changes. The quarterly posture update triggered by Microsoft's Agent 365 and Copilot capability releases is the operational mechanism. The Stage 3 maturity test in Section 10 is the quality gate. An organization that can pass that test for every agent currently in production is operating at the standard the regulatory environment is moving toward.

The Closing Observation

There is an enterprise pattern that repeats reliably in the adoption of any powerful new technology. The organization acquires the capability faster than it designs the governance. The governance gap accumulates invisibly. An incident or a regulator eventually surfaces it. The organization then spends significantly more time and money on retroactive governance than it would have spent building it correctly in the first place.

Agentic AI is running that pattern now, at speed. The agents are deployed. The governance architecture that should sit beneath them has not been built at the same pace.

What makes the current moment different from previous technology adoption cycles is that the regulatory language is moving faster than usual. FINRA's December 2025 guidance expects firms to maintain formal review and approval processes and comprehensive documentation throughout the AI lifecycle. The April 2026 interagency model risk revision expects clear roles and responsibilities with well-defined accountability. The Federal Reserve expects human accountability for decisions. These are not aspirational statements. They are examiner anchors. They describe the artifacts an examiner will ask for when they find an agent in production.

The Intent Architecture Stack is the organizational design framework that produces those artifacts before they are requested. Three layers, built before the agent goes live, updated when the platform changes, maintained until the agent is retired. When all three are present, the board question has an answer. When any of them are absent, the organization is one incident away from building them under pressure.

> **The platform can lock doors. It does not invent your org chart. That work belongs to the organization, and the organization that does it before the first incident will have a very different conversation with its board than the one that does it after.**

About This Paper and the Framework

This white paper presents original research and framework development by Sougata Roy, published at sougataroy.com and through The Governance Gap newsletter. The Intent Architecture Stack is version 1.0, dated May 2026. Every statistic cited is sourced from a named primary source with a publication date. Where a source could not be verified, it was excluded.

Research cutoff. This paper reflects regulatory guidance, litigation status, Microsoft product capabilities, and statistical reports as of May 26, 2026. Regulatory language, court docket activity, product capabilities, and industry statistics may have changed since that date. The framework constructs, diagnostic methodology, and organizational design principles are durable and apply regardless of subsequent regulatory development. Readers consulting this paper after May 2026 should verify current regulatory positions, product capabilities, and litigation status against primary sources before making deployment decisions.

Originality and attribution. The contribution of this paper is the integration of established and concurrent enterprise governance vocabulary into a coherent organizational design framework for agentic AI accountability in regulated enterprises. Three of these constructs are introduced as formal named constructs in this paper: the three-layer Intent Architecture Stack as an organizational design pattern, the Consequence Owner role construct, and the three-tier risk model for agent deployment. The remaining original constructs (the Governance Debt Maturity Model with its three stages of Accumulation, Recognition, and Resolution; the diagnostic instruments and intent documentation template; and the regulator-language mapping in Section 6) have their earliest documented public use at sougataroy.com framework pages and in The Governance Gap newsletter on LinkedIn, editions 1 through 11 (March 20 through May 26, 2026), with specific first-use dates noted in the per-term acknowledgments below. Several supporting terms used in this paper appear in established or concurrent industry and academic literature; each is acknowledged in the per-term notes that follow, with

the specific application to agentic AI accountability developed in this paper claimed as the distinctive contribution rather than the underlying term.

Intent Architecture. The phrase appears in prior literature in unrelated technical senses, most prominently in intent-based networking (Bezahaf et al., 2021; Alcock et al., 2022; Iordache et al., 2025), in human-agent teaming for command communication (Schneider and Miller, 2018), in software architecture patterns (Chauhan et al., 2021), in infrastructure-as-code (Allam, 2025), and in defense command systems (Lee et al., 2018). The use in this paper refers specifically to the organizational design layer for AI agent accountability and is distinct from these prior technical senses.

Governance Debt. The construct appears in two adjacent uses concurrent with or prior to the earliest public use of the term in this work. Nair (Validation Debt as a Source of Systemic Risk in Enterprise Software Systems, ResearchGate preprint, March 2026) develops Governance Debt as one dimension within a Validation Debt taxonomy in software validation, defined as the absence of ownership, measurement, accountability, and decision authority for validation effectiveness. Kachakayala (Building Secure Enterprise Data Lakes on Azure, International Journal of AI, BigData, Computational and Management Studies, February 2026) introduces Governance Debt in enterprise data lake architecture, defined as the deferred cost of failing to implement proper data controls, lineage, and definitions at project outset. Both works extend Cunningham's (1992) technical debt metaphor into organizational obligation territory. The use in this paper is to the agentic AI deployment accountability layer, paired with the three-stage Governance Debt Maturity Model. The shared metaphor reflects independent extensions across adjacent enterprise-technology domains; the unit being governed (software validation, data architecture, AI agent accountability) and the operational framework proposed for each differ accordingly.

Agent Sprawl. Established industry vocabulary predating March 20, 2026 for the uncontrolled proliferation of AI agents across an enterprise. The term appears in McKinsey (Seizing the Agentic AI Advantage, 2025), the Google AI Agents white paper (2025), the Berkeley California Management Review article by Saini (March 2026), and multiple academic preprints in 2026. The three-tier operationalization developed in Section 7 (employee shadow AI; organizational procurement without central visibility; authorized agents with over-permissive operational scope) is the distinctive contribution of this paper.

Intent Gap. Established AI and software engineering vocabulary since 2024 for the distance between user or developer intent and AI or program behavior. Prior uses include visual content generation (Cheng et al., 2024), intent formalization in software engineering (Lahiri, 2026), agentic program repair (Wang and Huang, 2026), and vision foundation models (Wu et al., 2025). The use in this paper extends the construct from specification-fidelity to organizational monitoring: the unplanned divergence between an organization's documented intent for an AI agent and the agent's actual behavior in production, framed as a governance monitoring failure rather than a specification or execution failure.

The Accountability Assumption. The phrase appears in prior literature in unrelated senses, including political theory on accountability in liberal democratic governance (Flinders, 2023, citing Dubnick), stakeholder theory and corporate social responsibility disclosure (Ullah et al., 2026), education policy on teacher preparation evaluation (Cochran-Smith and Reagan, 2022), and distributed systems theory on Byzantine consensus protocols. The use in this paper refers specifically to the implicit organizational belief that legal responsibility for an AI agent's consequential decisions resides with the vendor, the platform, or another team rather than with the deploying organization, a construct specific to agentic AI deployment in regulated enterprises.

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). Both are introduced in this paper.

Trademark verification. Exact-phrase searches across all seven named terms (Intent Architecture, Intent Architecture Stack, Governance Debt, Agent Sprawl, Intent Gap, Consequence Owner, The Accountability Assumption) through United States Patent and Trademark Office TESS in May 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 framework developed in this paper.

AI assistance. Drafting of this paper was developed with large language model assistance. The framework structure, the diagnostic methodology, the organizational design choices, and the conclusions are the original work of the author, refined through independent practitioner experience in regulated enterprise environments. All citations, incident facts, and regulatory references were verified against primary sources as of May

2026. The provenance of individual vocabulary terms is documented in the per-term notes above.

Citation format. Roy, S. (2026). *Who Owns the Agent? The Intent Architecture Stack Framework White Paper, Version 1.0.* sougataroy.com. May 2026. ORCID: https://orcid.org/0009-0002-9294-2566. DOI: 10.5281/zenodo.20481551.

Version history. Version 1.0, initial public release, May 2026. Research cutoff: May 26, 2026. Author ORCID: https://orcid.org/0009-0002-9294-2566. Subsequent versions, if issued, will be noted here with their release date and a summary of changes.

References

Aim Security (Aim Labs). (2025). EchoLeak: zero-click prompt injection in Microsoft 365 Copilot, disclosed June 2025. Tracked as CVE-2025-32711 (CVSS 9.3); patched by Microsoft in June 2025. National Vulnerability Database, nvd.nist.gov/vuln/detail/CVE-2025-32711. Accessed May 31, 2026.

Amstad, M. (2025). "Innovating Financial Supervision with SupTech." Singapore Fintech Festival, November 13, 2025.

British Columbia Civil Resolution Tribunal. (2024). *Moffatt v. Air Canada*, 2024 BCCRT 149. February 2024.

CISA, NSA, ASD ACSC, CCCS, NCSC-NZ, and NCSC-UK. (2026). Careful Adoption of Agentic AI Services. May 1, 2026. cisa.gov/resources-tools/resources/careful-adoption-agentic-ai-services

Cisco. (2026). State of AI Security 2026. Cisco AI research team, February 2026. cisco.com/site/us/en/products/security/state-of-ai-security.html. Accessed May 31, 2026.

Cloud Security Alliance. (2026). NIST AI RMF Agentic Profile. March 2026.

Di Castri, S., Grasser, M., and Barasa, M. (2025). State of SupTech Report 2025. Cambridge SupTech Lab (University of Cambridge) and Digital Transformation Solutions. December 2025. SSRN 5903962.

Dunn v. Upstart Holdings, Inc., No. 3:26-cv-02974 (N.D. Cal., complaint filed April 7, 2026, Donato, J.). Public PACER docket via CourtListener,

courtlistener.com/docket/73155267/dunn-v-upstart-holdings-inc. Status as of May 28, 2026; no ruling on the merits issued.

European Union. (2024). Regulation (EU) 2024/1689 (the AI Act). Article 26.

Federal Reserve, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation. (2026). Interagency Guidance on Model Risk Management. April 2026.

FINRA. (2025). 2026 Annual Regulatory Oversight Report. December 2025.

IBM. (2025). Cost of a Data Breach Report 2025. 20th annual edition, research conducted independently by the Ponemon Institute, sponsored and published by IBM, July 2025. ibm.com/reports/data-breach. Accessed May 31, 2026.

ISO/IEC. (2023). ISO/IEC 42001:2023, Artificial Intelligence Management System.

Kistler et al. v. Eightfold AI Inc., No. 3:26-cv-01768 (N.D. Cal.), removed March 2, 2026 from Superior Court of California, County of Contra Costa, No. C26-00214 (complaint filed January 20, 2026); assigned to Judge Yvonne Gonzalez Rogers. Public PACER docket via CourtListener, courtlistener.com/docket/72351430/kistler-v-eightfold-ai-inc. Defendant's motion to dismiss filed April 20, 2026, hearing set August 4, 2026; no ruling on the merits issued as of May 2026.

Microsoft. (2026). Cloud Adoption Framework: AI agent governance guidance. learn.microsoft.com.

Microsoft. (2026). *Detecting and mitigating common agent misconfigurations*: *Copilot Studio agent security, top 10 risks to detect and prevent.* Microsoft Security Blog, early 2026.

Microsoft. (2026). Entra Agent ID overview. learn.microsoft.com/en-us/entra/agent-id.

Microsoft. (2026). Purview AI documentation. learn.microsoft.com.

Microsoft. (2024). Responsible AI Standard v2.

NIST. (2023). AI Risk Management Framework (AI RMF 1.0).

Norval, C., Cobbe, J., and Singh, J. (2022). Accountability after the fact: AI systems and answerability. *Data & Policy*.

Orca Security. (2026). RoguePilot disclosure. February 2026.

Reco. (2025). 2025 State of Shadow AI Report. reco.ai/state-of-shadow-ai-report. Accessed May 31, 2026.

Saviynt. (2026). 2026 CISO AI Risk Report. Published by Saviynt with Cybersecurity Insiders, February 2026; survey of 235 CISOs, CIOs, and senior security leaders. saviynt.com/ciso-ai-risk-report-2026. Accessed May 31, 2026.

Team8. (2025). AI, Risk, and the Road Ahead: Key Findings from the 2025 CISO Village Survey. July 17, 2025; survey of over 110 CISOs. team8.vc/ciso-village-survey-2025. Accessed May 31, 2026.

Waller, C. (2026). "Operationalizing AI at the Federal Reserve." Federal Reserve Bank of Boston Technology-Enabled Disruption Conference, February 24, 2026. federalreserve.gov/newsevents/speech/waller20260224a.htm

Woodruff, J. (2026). The Agentic Trust Framework: Zero Trust Governance for AI Agents. Cloud Security Alliance blog, February 2, 2026. cloudsecurityalliance.org/blog/2026/02/02/the-agentic-trust-framework-zero-trust-governance-for-ai-agents.

sougataroy.com | The Governance Gap | May 2026 | Views are my own.