Agents

Part III: How to Govern an Agent

Regulation, Controls, Conformance, and Deployment

Jillian Bommarito · Daniel Martin Katz · Michael J Bommarito II

November 21, 2025

Working Draft Chapter

Version 0.1

This chapter is Part III of a three-part series from the textbook *Artificial Intelligence for Law and Finance*, currently under development. Individual chapters are being drafted and refined independently before integration into the complete book. Part I: *What is an Agent?* and Part II: *How to Build an Agent* are available separately.

You can find the most current copy of the textbook project here:
`https://github.com/mjbommar/ai-law-finance-book/`

Contents

How to Read This Chapter

This chapter translates the conceptual foundations from Part I into governance practice. Where Part I asked *what makes a system agentic?*, this chapter addresses *how do we govern agentic systems responsibly?* Our focus is on what chief risk officers, compliance officers, general counsel, and senior leadership need to approve, monitor, and continuously improve deployments in regulated domains.

Reading Paths by Audience.

- **Executives and Board Members**: Focus on Sections 1 (Introduction), 3 (Governance Stack), 5 (Accountability), and 7 (Conclusion). These sections provide strategic framing, regulatory context, organizational structure, and synthesis.
- **Compliance, Risk, and Legal Officers**: Read the entire chapter sequentially. Sections 3 through 4 provide detailed frameworks for regulatory alignment, control calibration, and operational implementation.
- **Practitioners and System Owners**: Emphasize Sections 2 (Dimensional Calibration), 4 (Implementation), and 6 (Examples). These sections operationalize governance through risk-based control selection, technical architecture, and worked examples.

Conceptual Framework. This chapter builds directly on Part I's analytical framework. Part I established a three-level hierarchy: *agents* (Level 1: minimal agency with Goal, Perception, Action), *agentic systems* (Level 2/3: operational readiness adding Iteration, Adaptation, Termination), and *AI agents* (agentic systems powered by AI/ML). The six properties that define agentic systems—**Goal**, **Perception**, **Action**, **Iteration**, **Adaptation**, **Termination** (GPA+IAT)—map systematically to governance requirements. For example, autonomy and actuation scope determine oversight intensity (human-in-the-loop vs. human-in-command); entity frame and persistence drive audit logging and records retention; goal dynamics influence escalation triggers and revalidation schedules. Part I provided the taxonomy; this chapter provides the governance logic for agentic systems specifically.

What This Chapter Is Not. This is not a step-by-step compliance manual, nor does it constitute legal advice. Regulatory requirements vary by jurisdiction, sector, and organizational context. Our goal is to equip you with conceptual tools—dimensional calibration, risk-based control selection, organizational accountability structures—that enable you to design governance proportionate to your risk profile. Consult qualified legal, compliance, and technical experts when implementing governance in your organization.

---

1 Introduction: The Governance Imperative

Software has always required governance. We audit code, review changes, test deployments, and maintain access controls. Yet the governance challenges posed by agentic systems differ in *kind*, not merely degree. Understanding this shift begins with recognizing what makes agents fundamentally different from the passive tools that dominate enterprise software today—and why those differences create accountability obligations that traditional governance structures were not designed to address.

1.1 From Tools to Agents: The Governance Shift

Most enterprise software operates as a passive tool: you invoke it, it executes a predetermined sequence, and it stops. A spreadsheet recalculates when you enter data. A database returns results when you query it. A compiler translates source code when you run it. These tools are **reactive**—they wait for explicit human commands, execute well-defined operations, and produce outputs that can be traced directly to inputs and logic paths.

Governance for passive tools focuses on *authorization* (who can invoke the tool), *configuration* (what parameters are allowed), and *validation* (does the output match expectations). When a spreadsheet miscalculates, we examine the formulas. When a database returns incorrect results, we inspect the query and schema. The causal chain from invocation to outcome is short, deterministic, and observable.

Agents introduce **Goal**, **Perception**, and **Action**—the GPA properties from Part I. Part I established a three-level hierarchy: *agents* (Level 1) possess these three minimal properties; *agentic systems* (Level 2/3) add three operational properties—Iteration, Adaptation, and Termination—required for production deployment. An agent is not merely invoked; it is assigned an objective. It does not passively wait for instructions; it perceives its environment, evaluates possible actions, and selects behaviors designed to advance its goal.

This autonomy creates three immediate accountability challenges:

1. **Purpose Drift**: A tool does what you tell it to do. An agent interprets what you *want* it to achieve. If the goal specification is ambiguous, incomplete, or misaligned with actual intent, the agent may pursue objectives you did not intend. Governance must verify goal alignment before

deployment and monitor for drift during operation.

2. **Perceptual Opacity**: Agents make decisions based on what they perceive. If perception is incomplete, biased, or adversarially manipulated, actions may be inappropriate even if the goal is well-specified. Unlike a passive tool whose inputs are explicit function parameters, an agent's perceptual inputs may include external data sources, sensor readings, or inferred environmental state. Governance must establish *input validation*, *data provenance*, and *bias detection* mechanisms.

3. **Actuation Risk**: Agents take actions that affect their environment—filing documents, executing trades, sending communications, modifying databases. Unlike passive tools that produce outputs for human review, agents *do things*. If an agent's action set includes high-consequence operations (e.g., signing contracts, disbursing funds, disclosing confidential information), governance must enforce *approval gates*, *actuation constraints*, and *rollback capabilities*.

Scope: This Chapter Focuses on Agentic Systems

This chapter addresses the specific governance challenges posed by **agentic systems**—AI systems exhibiting all six operational properties (Goal, Perception, Action, Iteration, Adaptation, Termination) as defined in Part I. While agentic systems present unique accountability and compliance challenges, they represent only one category within the broader landscape of AI technologies deployed in legal, financial, and audit contexts.

Many critical AI governance questions—such as foundation model evaluation, training data provenance, algorithmic fairness in non-agentic classifiers, explainability requirements for static models, and sector-specific ethical considerations—extend beyond agentic systems to AI more generally. These broader governance considerations, including frameworks for non-agentic AI applications, will be addressed in a forthcoming companion volume dedicated to comprehensive AI governance across all system types.

The GPA Governance Gap

Traditional software governance assumes human-in-the-loop execution: humans decide when to invoke tools, interpret outputs, and take consequential actions. GPA properties move decision-making *inside* the system boundary. Governance must shift from *access control* to *behavioral oversight*.

Note on Human Agents: Many of these governance controls—goal authorization, perceptual validation, actuation constraints—mirror requirements we impose on human agents (employees, contractors, delegates). When we hire a paralegal or junior analyst, we specify their objectives, verify the quality of their information sources, and limit their authority to take binding actions. The GPA framework makes explicit what has long been implicit in human delegation: *agency requires accountability structures*. What differs for AI agents is the need to encode these controls in technical systems rather than organizational policies alone.

If GPA creates accountability challenges for basic agents, the properties that define full **agentic systems—Iteration**, **Adaptation**, and **Termination** (IAT)—amplify them. Iteration means the system operates across multiple perceive-act cycles, each depending on prior state and environmental feedback. Governance must maintain *audit trails* that reconstruct decision sequences and enable reproducibility. Adaptation means the system changes its strategy based on experience; governance must implement *change control* and continuous revalidation. Termination means the system must know when to stop, hand off to a human, or escalate; governance must define *exit protocols* and *emergency stop mechanisms*.

These properties combine multiplicatively. An agentic system that adapts its perception across iterated interactions while pursuing evolving goals creates a governance surface far larger than a deterministic, single-invocation tool.

1.2 The Stakes: Professional Duties Are Non-Delegable

The governance imperative becomes urgent when we recognize a foundational legal and professional principle: **professional duties cannot be delegated to AI**. Attorneys, investment advisers, auditors, and other licensed professionals remain fully liable for the quality, accuracy, and ethical propriety of their work product—regardless of whether they used AI assistance.

Consider three domains:

Legal Practice. The American Bar Association's Model Rules of Professional Conduct impose duties of *competence* (Rule 1.1), *confidentiality* (Rule 1.6), and *candor to the tribunal* (Rule 3.3) on attorneys personally. When an attorney files a brief containing AI-generated citations, the attorney is responsible for verifying those citations exist and support the legal argument. In *Mata v. Avianca*, Inc., an attorney submitted a brief with hallucinated case citations generated by ChatGPT—a single-shot text generator lacking the iteration, tool access, and verification loops that would characterize an agentic legal research system (United States District Court Southern District of New York 2023). The court sanctioned the attorney—not the AI vendor—because the professional duty to verify legal research is non-delegable (American Bar Association Standing Committee on Ethics and Professional Responsibility 2024). This case illustrates that even non-agentic AI tools create professional responsibility obligations; agentic systems with autonomous iteration and actuation capabilities demand even greater governance.

Financial Services. Investment advisers owe fiduciary duties to clients under the Investment Advisers Act of 1940. This includes duties of care (providing suitable advice) and loyalty (acting in the client's best interest). If an adviser uses an AI chatbot to generate portfolio recommendations, the adviser remains liable for ensuring those recommendations are suitable, free from conflicts of interest, and supported by adequate analysis. "The AI recommended it" is not a defense to a breach of fiduciary duty claim.

Audit and Accounting. The Public Company Accounting Oversight Board (PCAOB) requires auditors to exercise *professional skepticism* and maintain *independence* when auditing financial statements. If an auditor uses AI to select samples for testing or analyze accounting estimates, the auditor must understand the tool's methodology, validate its outputs, and document the rationale in workpapers. The auditor cannot delegate professional judgment to the AI and remain compliant with PCAOB standards (Public Company Accounting Oversight Board 2010a; Public Company Accounting Oversight Board 2010b).

"The AI Did It" Is Not a Defense

Across legal, financial, and audit domains, professional responsibility rules establish that using AI tools does not diminish the professional's accountability. Governance is not optional—it is the operational mechanism for maintaining professional competence and fulfilling non-delegable duties.

1.3 Three Forces Driving Governance Adoption

Beyond professional obligations, three converging forces make governance essential for any organization deploying agentic systems:

Regulatory Momentum. AI-specific regulation is no longer hypothetical. The European Union's AI Act entered into force in August 2024, establishing risk-based requirements for high-risk AI systems including those used in credit decisioning, employment, law enforcement, and critical infrastructure (European Parliament and Council 2024). Systems classified as high-risk must undergo conformity assessments, maintain documentation, implement human oversight, and enable auditability—or face penalties up to €35 million or 7% of global annual turnover.

In the United States, sector-specific regulators are issuing guidance at an accelerating pace. The Federal Reserve's SR 11-7 guidance on model risk management applies to AI/ML systems used by banking institutions (Board of Governors of the Federal Reserve System 2011). The Equal Credit Opportunity Act requires lenders to provide "principal reasons" for adverse credit decisions, a requirement that extends to AI-driven underwriting (Consumer Financial Protection Bureau 2011). States are enacting their own requirements: Colorado's AI Act (effective January 2026) prohibits algorithmic discrimination and requires impact assessments for high-risk systems (Colorado General Assembly 2024).

This regulatory patchwork means organizations cannot rely on a single compliance framework.

Governance must layer multiple obligations.

Liability Exposure. Early litigation is establishing precedents that governance gaps create liability. *Mata v. Avianca* demonstrated that attorneys cannot blame AI for professional failures. Fair lending cases under the Equal Credit Opportunity Act impose strict liability—intent is not required; disparate impact suffices. If an AI credit scoring model produces outcomes that disproportionately harm protected classes, the lender is liable regardless of whether the model was "neutral" or purchased from a reputable vendor.

Vendor contracts typically shift risk to deployers through liability caps, warranty disclaimers, and indemnification clauses. A foundation model vendor may cap damages at the subscription fee—often insufficient to cover regulatory penalties, reputational harm, or class action settlements. Governance—demonstrating reasonable care through risk assessment, validation, monitoring, and incident response—becomes the primary defense.

Trust and Reputation. Legal, financial, and audit services are *trust-intensive* domains. Clients hire attorneys because they trust professional judgment. Investors entrust assets to advisers based on fiduciary obligations. Public companies rely on auditors to provide independent assurance. AI failures that compromise accuracy, confidentiality, or impartiality erode this trust irreparably.

A law firm that discloses client confidential information through an AI tool's training data breach faces not only regulatory sanctions but client defection. An investment adviser whose AI chatbot provides unsuitable recommendations faces not only fiduciary duty claims but loss of clients. An audit firm whose AI sampling tool produces biased or incomplete samples faces not only PCAOB sanctions but damage to its reputation for independence.

In trust-intensive domains, governance is not merely a compliance obligation—it is a competitive necessity.

1.4 Mapping Agent Properties to Governance Requirements

Effective governance begins with a systematic mapping from the technical properties that define agentic behavior (the GPA+IAT framework from Part I) to the specific controls required to manage risk, ensure compliance, and maintain accountability. This section provides that mapping, organized by property.

Note on System Architecture: This chapter assumes familiarity with the GPA+IAT framework from Part I. Organizations evaluating whether a specific system qualifies as an "agentic system" should apply Part I's six-question rubric and falsification tests. Part II (*How to Build an Agent*) covers specific architectures (ReAct, Reflexion, tool-calling frameworks) and helps teams distinguish agentic systems from sophisticated chatbots or single-shot inference systems.

Goal: Purpose Limitation and Alignment. An agent's goal determines what it optimizes for. Governance must ensure goals are:

- **Authorized**: Who specifies the goal? Under what authority? In regulated domains, goal-setting may require approval from compliance officers, general counsel, or clients.
- **Aligned**: Does the goal match actual organizational or client objectives? Misaligned goals—optimizing for throughput at the expense of quality, minimizing cost without considering risk—create liability.
- **Bounded**: What constraints limit goal pursuit? Agents may pursue goals too aggressively, ignoring side effects, ethical boundaries, or resource limits.
- **Monitorable**: Can we detect when the agent is not achieving its goal or when goal pursuit causes unintended harms? Governance must establish KPIs and SLAs that track goal satisfaction and side-effect metrics.

Perception: Data Governance and Input Validation. An agent's perception defines what information it uses to make decisions. Governance must address:

- **Provenance**: Where does the data come from? Is it authoritative, current, and trustworthy? Agents that perceive stale, fabricated, or biased data will make flawed decisions. For third-party systems, establishing provenance can be exceedingly difficult. Governance must include *vendor assessment protocols* and document provenance gaps as residual risk.
- **Bias and Representation**: Does the agent's perceptual model reflect population diversity, or does it encode historical biases? Governance must implement bias detection and fairness audits.
- **Input Validation**: Can adversaries manipulate what the agent perceives? Prompt injection, data poisoning, and adversarial examples are perceptual attacks. Governance must enforce input validation, sanitization, and anomaly detection.
- **Privacy and Confidentiality**: Does perception require access to sensitive data? Governance must ensure data minimization, encryption, and access controls that preserve confidentiality and comply with privacy regulations.

Action: Actuation Controls and Approval Gates. An agent's action set determines what it can *do*. Governance must manage actuation risk:

- **Action Authorization**: What actions is the agent permitted to take? High-consequence actions (e.g., signing contracts, disbursing funds) should require explicit authorization.
- **Pre-Action Approval**: Should certain actions require human approval before execution? Human-in-the-loop oversight is appropriate for irreversible or high-stakes actions.
- **Rollback and Remediation**: If an action causes harm, can it be undone? Governance must

design systems with rollback capabilities and remediation protocols.

- **Rate Limiting**: Can the agent take actions too quickly or too frequently? Governance must enforce rate limits and circuit breakers.

Iteration: State Management and Audit Trails. Iteration means the system operates across multiple cycles, each building on prior state. Governance must ensure:

- **Reproducibility**: Can we replay the system's decision sequence? Debugging, auditing, and compliance reviews require reconstructing what the system perceived and why it acted as it did.
- **State Integrity**: Is the system's internal state protected from tampering or corruption? Governance must implement tamper-evident logging and state validation.
- **Termination Conditions**: Does the system know when to stop iterating? Governance must define termination criteria (e.g., goal achieved, resource limit reached, safety violation detected).

Adaptation: Change Control and Revalidation. Adaptation means the system's behavior changes over time. Governance must manage behavioral drift:

- **Change Detection**: When did the system's behavior change? What triggered the adaptation? Governance must implement model versioning and change logs.
- **Revalidation Triggers**: Does adapted behavior still satisfy safety, fairness, and compliance constraints? Governance must define revalidation triggers (e.g., performance degradation, distribution shift, policy updates).
- **Rollback to Known-Good States**: If adaptation introduces failures, can we revert to a prior validated version?
- **Human Oversight of Learning**: Should adaptation require human approval? In high-stakes domains, unsupervised learning may be inappropriate.

Termination: Exit Protocols and Escalation. Termination governs when and how the system stops operating. Governance must define:

- **Escalation Triggers**: Under what conditions does the system hand off to a human? Examples include ambiguous inputs, conflicting objectives, safety violations, or low-confidence decisions.
- **Graceful Shutdown**: How does the system cleanly exit? Abrupt termination may leave systems in inconsistent states.
- **Handoff Procedures**: When the system escalates to a human, what information must it provide? Effective handoff requires context.
- **Override and Emergency Stop**: Can humans immediately halt the system? Governance must

provide emergency stop mechanisms (the "red button") accessible to authorized personnel.

From Properties to Controls

The GPA+IAT framework is not merely a taxonomy for understanding agents—it is a *requirements map* for governance. Each property creates specific risks; each risk demands specific controls. Organizations that deploy agentic systems without systematically addressing all six properties face gaps in accountability, compliance, and safety.

The remainder of this chapter builds on this foundation. Section 2 shows how to calibrate control intensity based on system autonomy, entity frame, goal dynamics, and persistence—establishing the control logic that governs agentic systems. Section 3 then maps regulatory obligations into a five-layer framework, demonstrating how to apply these calibrated controls across regulatory layers. Sections 4 and 5 translate principles into operational practices and organizational structures. Section 6 demonstrates governance through worked examples in legal, financial, and audit contexts. Section 7 synthesizes the governance imperative and provides a maturity-based path forward.

---

2 Design Principles: Dimensional Calibration

This section establishes *how much* control intensity is required—how to calibrate governance based on system properties. Section 3 will then demonstrate *what* controls are required across regulatory layers, applying this calibration framework to specific legal and professional obligations. The key insight: governance is not binary (present or absent) but dimensional (scaled to risk). Organizations that apply uniform controls to all agentic systems either over-engineer low-risk deployments (wasting resources) or under-protect high-risk deployments (creating liability exposure). Dimensional calibration matches governance intensity to the operational characteristics that drive risk.

2.1 The Dimensional Calibration Logic

Before applying dimensional calibration, verify the system qualifies as an *agentic system* under Part I's framework:

GPA+IAT Validation Requirement

Dimensional calibration applies **only to agentic systems**—those meeting all six operational properties:

GPA (Level 1 Agent):

- **Goal**: Clear objective directing behavior?
- **Perception**: Observes environment/inputs?
- **Action**: Affects environment/produces outputs?

+IAT (Agentic System):

- **Iteration**: Multiple perception-action cycles?
- **Adaptation**: Modifies strategy based on feedback?
- **Termination**: Explicit or implicit stopping conditions?

Systems lacking any of these six properties are *not agentic systems*. A single-shot ML model, batch classifier, or non-iterative chatbot requires different governance approaches beyond this chapter's scope. **Verify GPA+IAT compliance before proceeding to dimensional calibration.**

We calibrate governance intensity across four analytical dimensions introduced in Part I:

1. **Autonomy**: The degree of independence the system exercises in decision-making. Ranges from human-in-the-loop (HITL) requiring pre-approval for every significant action, through human-on-the-loop (HOTL) where humans monitor and can intervene, to human-in-command (HIC) where humans set strategic goals and retain emergency stop authority but do not review individual decisions.
2. **Entity Frame**: How the system presents itself and how users perceive its role. Ranges from *human frame* (agent represents a specific professional), through *hybrid* (collaborative partnership), to *machine frame* (clearly identified as non-human tool), to *institutional frame* (agent acts on behalf of the organization).
3. **Goal Dynamics**: How the system's objectives change over time. Ranges from *static* (fixed goals validated once), through *adaptive* (system refines goals within predefined boundaries based on feedback), to *negotiated* (system proposes goal changes requiring explicit human approval).
4. **Persistence**: Whether the system maintains state across interactions. Ranges from *stateless* (each interaction independent) to *stateful* (system accumulates information, builds context, and decisions depend on interaction history).

Risk is **multidimensional**, not unidimensional. A system's overall risk profile emerges from the *combination* of autonomy, entity frame, goal dynamics, and persistence. Control intensity must respond to this combination. The following subsections calibrate each dimension independently, then Section 2.6 demonstrates integration.

Why Dimensional Calibration Matters

Generic governance frameworks provide one-size-fits-all guidance: "implement human oversight," "maintain logs," "ensure fairness." Dimensional calibration operationalizes these principles: *how much* human oversight (HITL vs. HOTL vs. HIC)? *how detailed* must logs be (decision rationale vs. inputs/outputs only)? *how frequently* must fairness be validated (pre-deployment only vs. continuous monitoring)?

Without calibration, organizations default to either maximum controls (expensive, slow, may not be technically feasible) or minimum controls (cheap, fast, exposes liability). Calibration enables proportionate governance: controls sufficient to manage risk without unnecessary overhead.

2.2 Autonomy Calibration

Autonomy determines the degree of human involvement in decision-making. We distinguish three oversight modes, ordered by increasing system autonomy:

Human-in-the-Loop (HITL): Pre-Approval Required. In HITL mode, the system recommends actions but a human must approve before execution. HITL is appropriate when:

- Actions are high-consequence and irreversible (e.g., filing court documents, executing large trades, signing contracts).
- Professional duties require human judgment (e.g., attorney competence obligations, fiduciary duty).
- Regulatory requirements mandate human review (e.g., certain medical diagnoses, credit decisions in some jurisdictions).

Governance implications: HITL systems require approval workflows, notification mechanisms, and clear assignment of approval authority. Logging must capture both the system's recommendation and the human's decision (approve, reject, modify). Because humans review every significant action, post-action monitoring can be lighter—the human reviewer serves as the primary control. However, organizations must guard against automation bias: humans rubber-stamping AI recommendations without meaningful review.

Example: A legal research assistant that *iteratively* searches case law: it queries legal databases, evaluates result relevance, refines search terms based on findings, and generates progressive case summaries—all before presenting final output to an attorney for review. The attorney verifies citations, assesses legal reasoning, and takes responsibility for the final work product. **Note**: If the attorney must approve each individual search query before the next query executes, the system lacks autonomous iteration and is not a full agentic system despite having other properties.

Human-on-the-Loop (HOTL): Monitoring with Intervention. In HOTL mode, the system operates autonomously within defined parameters, but humans monitor performance and can intervene if anomalies, errors, or safety concerns arise. HOTL is appropriate when:

- Actions are moderate-consequence or reversible (e.g., customer service responses, preliminary data analysis).
- Real-time human review would create unacceptable latency but oversight remains necessary.
- The system operates within well-defined boundaries (e.g., credit limits, risk parameters).

Governance implications: HOTL systems require monitoring dashboards, anomaly detection, escalation triggers, and intervention protocols. Logging must be sufficiently detailed to enable retrospective review—since humans do not approve every action prospectively, they must be able to audit decisions retrospectively. Escalation triggers define when the system must halt and request human guidance (e.g., low-confidence decisions, outcomes near policy boundaries, user complaints).

Example: A customer service chatbot that handles routine inquiries autonomously but escalates complex questions, complaints, or regulatory issues to human agents. Supervisors monitor conversation logs, error rates, and escalation frequency.

Human-in-Command (HIC): Strategic Oversight and Emergency Stop. In HIC mode, the system operates with high autonomy. Humans set strategic goals, define constraints, and monitor aggregate performance but do not review individual decisions. Humans retain emergency stop authority to halt the system if safety violations, systemic failures, or regulatory concerns emerge. HIC is appropriate when:

- The system operates at scale and speed that precludes individual review (e.g., fraud detection processing millions of transactions daily).
- Actions are individually low-consequence but cumulatively significant.
- The system operates in a stable, well-understood environment with strong safeguards.

Governance implications: HIC systems require exceptionally strong logging, monitoring, and retrospective audit capabilities. Because humans do not review decisions prospectively or monitor continuously at the individual level, post-action auditability becomes critical. Organizations must implement statistical monitoring (e.g., fairness metrics, error rate trends, drift detection) to identify systemic issues. Emergency stop mechanisms must be accessible to authorized personnel and tested regularly.

Example: A fraud detection system that automatically blocks transactions meeting defined risk criteria. Fraud analysts set risk parameters, monitor aggregate block rates and false positive rates, and investigate flagged cases retrospectively. The system can be halted immediately if systemic bias or operational failures are detected.

The Autonomy-Auditability Trade-off. As autonomy increases, the burden of governance shifts from ex-ante (pre-approval) to ex-post (logging, monitoring, audit). HITL systems rely on human review as the primary control; HIC systems rely on comprehensive logging and statistical monitoring. Organizations must invest in monitoring infrastructure proportionate to autonomy: high-autonomy systems cannot rely on "we'll review it if someone complains."

Table 1 summarizes autonomy calibration.

Table 1: Autonomy Calibration: Oversight Modes and Control Requirements

Autonomy Level	Description	Example Use Cases	Control Requirements
HITL (Human-in-the-Loop)	Human pre-approves significant actions	Legal research, investment advice, contract review	Approval workflows, automation bias mitigation, competence training
HOTL (Human-on-the-Loop)	System operates autonomously; humans monitor and intervene	Customer service chatbots, preliminary audit analytics	Monitoring dashboards, escalation triggers, intervention protocols
HIC (Human-in-Command)	High autonomy with strategic oversight and emergency stop	Fraud detection, credit pre-screening (within parameters), algorithmic trading	Comprehensive logging, statistical monitoring, emergency stop, fairness metrics

2.3 Entity Frame Calibration

Entity frame determines how the system presents itself and how users perceive its role. Entity frame affects trust, liability allocation, and user expectations. Mismatches between entity frame and governance create risk.

Human Entity Frame. The system represents a specific human professional (e.g., "your attorney," "your financial adviser"). Users may not distinguish between the professional and the AI tool.

Governance implications: Human frame creates the highest accountability expectations. Professional responsibility rules apply in full. The professional represented by the system bears liability for all outputs. Confidentiality, competence, and fiduciary duty obligations are non-delegable. Governance must ensure the professional reviews, validates, and takes ownership of AI-generated outputs.

Mismatch risk: If the system operates with high autonomy (HIC) but presents a human frame, users may assume human oversight that does not exist. This creates misplaced trust and potential liability.

Example: A legal research tool that produces work product under the attorney's name. The attorney must verify citations, assess legal reasoning, and ensure compliance with Rule 1.1 (competence) and Rule 3.3 (candor).

Hybrid Entity Frame. The system is presented as a collaborative partnership between human and AI (e.g., "AI-assisted analysis," "our team uses advanced tools").

Governance implications: Hybrid frame requires clear delineation of responsibilities. Users should understand that AI provides preliminary analysis or recommendations, but humans make final decisions. Transparency about the division of labor reduces misplaced trust. Governance must document which tasks are AI-performed vs. human-performed and ensure human review of AI outputs before client-facing use.

Example: An investment advisory firm that discloses: "Our financial plans combine AI-driven market analysis with our advisers' professional judgment and knowledge of your personal circumstances."

Machine Entity Frame. The system is clearly identified as a non-human tool (e.g., "AI chatbot," "automated system"). Users understand they are interacting with technology, not a human.

Governance implications: Machine frame sets appropriate expectations. Users are less likely to assume human judgment, empathy, or professional accountability. However, organizations must ensure the system's capabilities match user expectations—a chatbot labeled as "informational only" should not provide advice that creates reliance. Governance must include clear disclaimers, capability limitations, and escalation to humans for complex or high-stakes issues.

Example: A customer service chatbot that states: "I'm an AI assistant. I can help with account questions, but for disputes or complex issues, I'll connect you with a human agent."

Institutional Entity Frame. The system acts on behalf of the organization (e.g., "XYZ Bank's credit decisioning system," "our firm's compliance review tool"). The organization, not an individual, bears accountability.

Governance implications: Institutional frame allocates liability to the organization. This is appropriate for systems used in institutional decision-making (credit underwriting, hiring, fraud detection). Governance must include organizational oversight (board and executive accountability), institutional policies (acceptable use, risk appetite), and enterprise-level monitoring. Professional responsibility considerations (if applicable) must be addressed separately.

Mismatch risk: If an institutional system operates without adequate organizational oversight (e.g., deployed by a rogue team without executive approval), the organization may face liability for decisions it did not authorize.

Example: A bank's credit pre-screening system that evaluates mortgage applications under institutional policies, with oversight by the Chief Risk Officer and compliance with ECOA.

Table 2 summarizes entity frame calibration.

Table 2: Entity Frame Calibration: Presentation Modes and Accountability Structures

Entity Frame	Description	Example Use Cases	Accountability Structure
Human	Agent represents specific professional	Legal research under attorney name, personalized financial advice	Professional bears full liability; professional responsibility rules apply; non-delegable duties
Hybrid	Collaborative human-AI partnership	AI-assisted audit analytics, co-drafted documents	Shared responsibility; clear delineation required; human validates AI outputs
Machine	Clearly identified as non-human tool	Customer service chatbot with AI disclosure, informational tools	Organization responsible for tool fitness; clear disclaimers and capability limitations
Institutional	Agent acts on behalf of organization	Credit decisioning, hiring, compliance review	Organizational liability; board/executive oversight; institutional policies and monitoring

2.4 Goal Dynamics Calibration

Goal dynamics determine how the system's objectives change over time. Static goals are easiest to govern; negotiated goals create the highest misalignment risk.

Static Goals. The system pursues a fixed objective defined at deployment. The goal does not change without explicit redeployment.

Governance implications: Static goals can be validated once during pre-deployment review. Organizations assess whether the goal aligns with organizational objectives, legal requirements, and ethical constraints. Once validated, the goal remains stable. Governance focuses on monitoring whether the system achieves the goal and whether side effects emerge.

Example: A legal research tool with the static goal: "Identify cases cited in the brief and verify they exist in official reporters." The goal does not change; the tool performs the same validation task repeatedly.

Adaptive Goals. The system refines its objectives within predefined boundaries based on feedback, but cannot change goals fundamentally. For example, a fraud detection system might adjust risk weights based on observed fraud patterns, but cannot change its core objective (detect fraud) or operate outside defined risk parameters.

Governance implications: Adaptive goals require continuous monitoring to ensure the system remains within boundaries. Organizations must define:

- **Boundaries**: What aspects of the goal can adapt? What constraints are inviolable?
- **Monitoring**: How frequently are adaptations reviewed? What triggers revalidation?
- **Rollback**: If adaptation degrades performance or violates constraints, can the system revert to a prior known-good state?

Example: A credit scoring model that adapts feature weights based on performance feedback but cannot introduce new features, change fairness constraints, or operate outside regulatory compliance boundaries.

Negotiated Goals. The system proposes changes to its objectives and requests human approval before implementation. This is the highest governance burden because each goal change requires validation.

Governance implications: Negotiated goals require human-in-the-loop approval for every proposed change. Organizations must establish:

- **Approval Authority**: Who can approve goal changes? (Typically requires senior leadership or governance committee.)
- **Change Justification**: Why is the system proposing the change? What evidence supports it?
- **Impact Assessment**: What are the consequences of the proposed goal change?
- **Revalidation**: After goal change, the system must be revalidated for safety, fairness, and compliance.

Example: An AI strategic planning assistant that proposes: "Based on market analysis, I recommend shifting investment focus from Technology to Healthcare." This goal change requires executive approval, risk assessment, and fiduciary duty review.

Table 3 summarizes goal dynamics calibration.

2.5 Persistence Calibration

Persistence determines whether the system maintains state across interactions. Stateful systems create compounding error risk and require state integrity controls.

Stateless Systems. Each interaction is independent. The system does not retain information from prior interactions.

Governance implications: Stateless systems are simpler to govern. Errors do not compound—a

Table 3: Goal Dynamics Calibration: Objective Stability and Governance Requirements

Goal Dynamics	Description	Example Use Cases	Governance Requirements
Static	Fixed objectives; no goal changes	Citation verification, rule-based compliance checks	One-time goal validation; monitor achievement and side effects
Adaptive	Refinement within boundaries based on feedback	Credit scoring (adjust weights), fraud detection (adapt risk parameters)	Define boundaries, continuous monitoring, rollback capability, revalidation triggers
Negotiated	System proposes goal changes requiring human approval	Strategic planning assistant, adaptive investment strategy	Approval workflows, impact assessment, revalidation after changes, senior leadership involvement

mistake in one interaction does not affect subsequent interactions. Logging can be lighter (capture inputs/outputs without state reconstruction). Reproducibility requires only input data, not interaction history.

Example: A legal citation verification tool that checks each citation independently. An error in verifying Citation A does not affect the verification of Citation B.

Stateful Systems. The system accumulates information across interactions. Decisions depend on prior state.

Governance implications: Stateful systems require state management controls:

- **State Integrity**: Protect state from tampering, corruption, or adversarial manipulation.
- **State Logging**: Capture state changes to enable decision reconstruction.
- **Error Compounding**: Monitor for cases where an initial error propagates through subsequent decisions.
- **State Reset**: Define conditions under which state should be reset (e.g., user logout, policy change, detected anomaly).

Example: A financial planning chatbot that builds a profile of the client's financial situation across multiple conversations. If the system misunderstands the client's risk tolerance in Session 1, all subsequent recommendations may be inappropriate. Governance must include periodic state validation ("Let me confirm: your risk tolerance is Moderate, correct?") and state logging to reconstruct how the profile evolved.

Table 4 summarizes persistence calibration.

Table 4: Persistence Calibration: State Management and Control Requirements

Persistence	Description	Example Use Cases	Governance Requirements
Stateless	Each interaction independent; no retained state	Citation verification, single-query research, one-time calculations	Standard logging (inputs/outputs); no state management; errors do not compound
Stateful	System maintains state across interactions; decisions depend on history	Multi-session financial planning, ongoing fraud monitoring, adaptive customer profiles	State integrity protection, state change logging, error compounding monitoring, periodic state validation

2.6 Integration: Risk-Calibrated Control Selection

Dimensional calibration becomes powerful when dimensions are integrated. A system's overall risk profile emerges from the *combination* of autonomy, entity frame, goal dynamics, and persistence. Controls must respond to this multidimensional risk.

Low-Risk Profile Example: Legal Research Assistant (HITL + Human + Static + Stateless).

- **Autonomy**: HITL—attorney must review and verify all outputs before use.
- **Entity Frame**: Human—tool operates under attorney's name and professional responsibility.
- **Goal Dynamics**: Static—fixed goal (verify citations exist and support arguments).
- **Persistence**: Stateless—each citation verification independent.

Control Calibration:

- **Logging**: Basic inputs/outputs sufficient (attorney review serves as primary control).
- **Monitoring**: Minimal (spot-check error rates quarterly).
- **Explainability**: Not required (attorney independently verifies legal reasoning).
- **Fairness**: Not applicable (legal research is not a protected-class decision).
- **Vendor management**: Standard SaaS due diligence (confidentiality, data retention).

This is a *low-governance* system because human review (HITL) compensates for other dimensions.

High-Risk Profile Example: Credit Underwriting (HIC + Institutional + Adaptive + Stateful).

- **Autonomy**: HIC—system makes credit decisions autonomously within risk parameters; humans monitor aggregate performance.
- **Entity Frame**: Institutional—system acts on behalf of the bank; organizational liability.
- **Goal Dynamics**: Adaptive—model adjusts feature weights based on performance feedback.
- **Persistence**: Stateful—system maintains applicant history across multiple applications.

Control Calibration:

- **Logging**: Comprehensive—capture inputs, model version, feature weights, decision rationale, human interventions, state changes. Retention: 25 months (ECOA) + 7 years (litigation hold).
- **Monitoring**: Continuous—monthly fairness metrics (80% rule, disparate impact ratios), data drift detection, approval rate trends by protected class, adverse action reason distributions.
- **Explainability**: ECOA-compliant—generate "principal reasons" for adverse decisions; validate explanations for faithfulness and completeness.
- **Fairness**: Pre-deployment validation + continuous monitoring + revalidation after model updates or detected distribution shift.
- **Vendor management**: Enhanced due diligence—model interpretability, fairness validation methodology, update notification, audit rights.
- **Human oversight**: Monthly compliance review, quarterly model performance review, board-level annual review.
- **Incident response**: Immediate halt if fairness violations detected; root cause analysis; regulator notification if systemic bias found.

This is a *high-governance* system because all four dimensions create compounding risk. High autonomy demands strong logging and monitoring. Institutional frame creates organizational liability. Adaptive goals require continuous revalidation. Persistent state creates error compounding risk. The combination necessitates intensive controls.

Dimensional Calibration Worksheet

When evaluating a new agentic system, assess each dimension:

1. **Autonomy**: HITL, HOTL, or HIC?
2. **Entity Frame**: Human, Hybrid, Machine, or Institutional?
3. **Goal Dynamics**: Static, Adaptive, or Negotiated?

4. **Persistence**: Stateless or Stateful?

Use Tables 1 through 4 to identify baseline controls for each dimension. Then integrate:

- High autonomy + institutional frame → Strong logging, statistical monitoring, board oversight.
- Adaptive goals + stateful persistence → Continuous revalidation, state integrity controls.
- HITL + human frame → Professional responsibility compliance, automation bias mitigation.

Dimensional calibration is not a formula—it is a structured reasoning framework that prevents under-protection ("it's just a chatbot") and over-engineering ("we must apply maximum controls to everything").

Section 4 operationalizes dimensional calibration through technical architecture and organizational processes. Section 6 demonstrates calibration through worked examples in legal, financial, and audit contexts.

---

3 The Governance Stack: Overlapping Obligations

Organizations deploying agentic systems in regulated domains face a complex, overlapping web of legal and professional obligations. There is no single "AI governance law" that comprehensively addresses all requirements. Instead, governance emerges from the interaction of five layers: foundational law, professional and ethical obligations, sector-specific regulation, AI-specific regulation, and voluntary governance frameworks. Understanding this layered structure—and why no single layer suffices—is essential for designing governance proportionate to organizational risk.

3.1 The Five-Layer Framework

We organize governance obligations into five layers, each building on the foundation below:

1. **Foundational Law**: Broadly applicable statutes governing data protection, discrimination, consumer protection, and contracts. Examples include the General Data Protection Regulation (GDPR), Equal Credit Opportunity Act (ECOA), and state consumer protection statutes. These laws establish baselines that apply regardless of whether AI is involved.
2. **Professional and Ethical Obligations**: Duties imposed on licensed professionals—attorneys,

investment advisers, auditors, accountants—by their governing bodies, bar associations, or regulatory agencies. These obligations are often more stringent than general law and impose fiduciary duties, confidentiality requirements, and competence standards.

3. **Sector-Specific Regulation**: Rules tailored to particular industries—banking, securities, insurance, healthcare—that address operational risks, supervision, and consumer protection in those domains. Examples include Federal Reserve guidance on model risk management (SR 11-7), Financial Industry Regulatory Authority (FINRA) rules on automated systems, and Public Company Accounting Oversight Board (PCAOB) auditing standards.

4. **AI-Specific Regulation**: Laws and regulations explicitly targeting artificial intelligence systems. The European Union's AI Act is the most comprehensive example, establishing risk-based requirements for high-risk AI systems. U.S. states (Colorado, California, New York City) are enacting their own AI-specific rules addressing bias, transparency, and impact assessments.

5. **Voluntary Governance Frameworks**: Standards, best practices, and certification schemes developed by standards bodies, industry groups, or government agencies. Examples include the NIST AI Risk Management Framework, ISO/IEC 42001 (AI Management Systems), COBIT for IT governance, and SOC 2 for vendor assurance. These frameworks are typically voluntary unless incorporated by reference into contracts or regulatory requirements.

Why Layering is Necessary

No single framework fully satisfies all governance requirements. The EU AI Act establishes high-level risk categories but does not specify how to comply with ECOA's "principal reasons" standard for adverse credit decisions. NIST AI RMF provides flexible risk management guidance but does not address attorney-client privilege or auditor independence. Organizations must layer multiple frameworks, augment them with domain-specific controls, and continuously monitor regulatory developments across jurisdictions.

The remainder of this section examines each layer in detail, identifying key requirements and illustrating how obligations interact.

3.2 Layer 1: Foundational Law

Foundational law provides the baseline for governance, applicable to all organizations regardless of industry or use case. Three domains are especially relevant:

Data Protection and Privacy: GDPR Article 22 and Stateful Agentic Systems. The GDPR establishes rights and obligations for processing personal data of EU residents. Article 22 addresses automated decision-making: individuals have the right not to be subject to decisions based solely on automated processing that produce legal or similarly significant effects (European Parliament and Council 2016). While not an absolute prohibition—automated decisions are permitted with explicit consent, contractual necessity, or legal authorization—Article 22 requires organizations to implement "suitable measures" to safeguard the data subject's rights, including the right to obtain human intervention and contest the decision.

Agentic-Specific Challenge—Stateful Decision Accumulation: Article 22's human intervention requirement becomes complex for stateful agentic systems that accumulate context across multiple cycles. Generic AI guidance suggests "add a button for human review," but this is insufficient for agentic systems. Meaningful human intervention requires access to the system's accumulated internal state—how the agent's understanding evolved across iterations, what adaptations occurred, what termination logic triggered the final decision. Without comprehensive state logging (capturing perception, action, and rationale at each cycle), human reviewers cannot meaningfully intervene or contest decisions because they lack visibility into how the agent reached its conclusion.

Governance Implication: For agentic systems subject to GDPR Article 22, human intervention controls must be paired with state logging requirements (cross-cycle audit trails). Organizations cannot satisfy Article 22 by providing post-hoc review without the ability to reconstruct the agent's iterative decision process. This links directly to the State Logging control discussed in Section 4.2.

Article 32 requires appropriate technical and organizational security measures, including encryption, pseudonymization, and resilience against unauthorized processing. Articles 33-34 mandate breach notification to supervisory authorities (within 72 hours) and affected individuals (without undue delay) when breaches pose risks to rights and freedoms.

For organizations operating globally, GDPR compliance often sets the de facto standard. Even organizations without EU operations may face GDPR obligations if they offer services to EU residents or monitor their behavior.

Anti-Discrimination and Fair Lending: ECOA and Process-Based Discrimination. The Equal Credit Opportunity Act (ECOA) prohibits credit discrimination based on race, color, religion, national origin, sex, marital status, age, or receipt of public assistance (United States Congress 1974). Regulation B, which implements ECOA, requires creditors to provide "principal reasons" for adverse credit decisions (Consumer Financial Protection Bureau 2011). This explainability requirement is more specific than generic AI transparency guidance: applicants must receive concrete, understandable reasons tied to the factors that most significantly influenced the decision.

ECOA applies regardless of whether the decision was made by a human, an algorithm, or a hybrid system. Courts have applied *disparate impact* theory: even facially neutral criteria can violate ECOA if they disproportionately harm protected classes without adequate business justification.

Agentic-Specific Challenge—Process-Based Discrimination: Traditional fairness testing focuses on *outcome parity* (do protected classes receive approvals at comparable rates?). For agentic credit underwriting systems that iteratively investigate applications across multiple cycles, discrimination can emerge through the *investigation process* itself, not just final decisions. An agentic system might:

- Adapt to request more verification cycles from applicants with characteristics correlated with protected classes (e.g., shorter U.S. employment tenure as a proxy for national origin).
- Impose higher process burdens (more documentation requests, longer investigation timelines) on protected groups, causing application abandonment even if the system would ultimately approve.
- Learn patterns that create disparate *iteration counts* across demographic groups, violating ECOA even when final approval rates satisfy the 80% rule.

Governance Implication: Agentic credit systems require *process parity monitoring* in addition to outcome fairness testing. Organizations must audit:

- Average investigation cycles by protected class (flag if deviation >20%).
- Termination reasons by demographic group (ensure similar rates of confidence-based vs. timeout-based termination).
- Application abandonment rates during multi-cycle investigation (ensure verification burdens do not disproportionately affect protected classes).

Governance must ensure the system's dynamic investigation strategy does not introduce prohibited discrimination—a challenge that does not arise with single-shot credit scoring models. See Section 4.6 for a worked example of process-based discrimination detected in agentic underwriting.

Consumer Protection. State consumer protection statutes (e.g., Unfair and Deceptive Acts and Practices laws) prohibit misleading representations and unfair business practices. If an AI chatbot misrepresents its capabilities, provides inaccurate information, or fails to disclose material limitations, the organization may face consumer protection enforcement regardless of whether the misrepresentation was intentional or the result of a model hallucination.

3.3 Layer 2: Professional and Ethical Obligations

Licensed professionals face heightened obligations that governance systems must operationalize. We examine three domains:

Legal Practice: ABA Model Rules. Most U.S. jurisdictions have adopted versions of the American Bar Association's Model Rules of Professional Conduct. Three rules are especially salient for AI governance:

- **Rule 1.1 (Competence)**: An attorney must provide competent representation, requiring "the legal knowledge, skill, thoroughness and preparation reasonably necessary for the representation." ABA Formal Opinion 512 (July 2024) clarifies that competence includes understanding

AI tools' capabilities and limitations (American Bar Association Standing Committee on Ethics and Professional Responsibility 2024). Attorneys cannot delegate legal analysis to AI without independent verification.

- **Rule 1.6 (Confidentiality)**: An attorney must not reveal information relating to representation of a client unless the client gives informed consent. Using AI tools that transmit client data to third-party vendors, train models on client information, or store data insecurely may violate confidentiality obligations. Governance must assess vendor data handling practices and obtain client consent where necessary.
- **Rule 3.3 (Candor Toward the Tribunal)**: An attorney must not knowingly make false statements of fact or law to a tribunal. Submitting AI-generated legal research without verification—resulting in fabricated citations or misrepresented holdings—violates this duty. The *Mata v. Avianca* case demonstrated that "the AI made the mistake" does not excuse the attorney's failure to verify (United States District Court Southern District of New York 2023).

Rule 5.3 addresses supervision of nonlawyer assistants, a framework some jurisdictions apply to AI tools. The attorney remains responsible for ensuring that AI-assisted work product meets professional standards.

Financial Services: Fiduciary Duty and Suitability. Investment advisers owe fiduciary duties to clients under the Investment Advisers Act of 1940 and SEC interpretations. This duty has two components:

- **Duty of Care**: Providing advice that is suitable for the client's financial situation, investment objectives, and risk tolerance. An AI-generated portfolio recommendation must be validated against these client-specific factors.
- **Duty of Loyalty**: Acting in the client's best interest, including full disclosure of conflicts of interest. If an AI tool is provided by an affiliate, receives compensation from recommended products, or prioritizes firm profitability over client outcomes, the adviser must disclose these conflicts and ensure recommendations remain in the client's best interest.

FINRA (Financial Industry Regulatory Authority) Rule 2111 imposes a similar suitability obligation on broker-dealers. Rule 3110 requires firms to supervise associated persons and establish procedures to ensure compliance. For AI systems that generate investment recommendations or execute trades, governance must include supervisory review procedures, monitoring for suitability violations, and escalation protocols.

Audit and Accounting: Independence and Professional Skepticism. The AICPA Code of Professional Conduct and PCAOB auditing standards impose strict independence and competence requirements on auditors:

- **Independence**: Auditors must maintain both independence in fact and independence in appearance. If an AI tool is provided by the audit client, an affiliate, or a vendor with financial ties to the client, independence may be impaired. The SEC and PCAOB closely scrutinize auditor-provided tools that could create management decision-making or self-review threats.
- **Professional Skepticism**: PCAOB Auditing Standard 1015 requires auditors to exercise professional skepticism—a questioning mind and critical assessment of audit evidence. Auditors cannot accept AI outputs uncritically; they must understand the tool's methodology, validate its logic, and assess whether results are consistent with other evidence.
- **Documentation**: AS 1215 (Audit Documentation) requires auditors to document the nature, timing, extent, and results of audit procedures. If AI is used for sampling, risk assessment, or analytical procedures, the workpapers must explain the tool's logic, parameters, and the auditor's rationale for relying on its output (Public Company Accounting Oversight Board 2010c).

These professional obligations are non-delegable. Governance systems must operationalize competence, confidentiality, independence, and documentation requirements through technical controls and organizational processes.

3.4 Layer 3: Sector-Specific Regulation

Sector regulators impose industry-tailored requirements that general frameworks do not address:

Banking: Model Risk Management. The Federal Reserve, Office of the Comptroller of the Currency (OCC), and Federal Deposit Insurance Corporation (FDIC) issued joint guidance SR 11-7 on model risk management (Board of Governors of the Federal Reserve System 2011). SR 11-7 applies broadly to model risk management in banking. Key requirements include:

- **Model Inventory**: Maintain a comprehensive inventory of models, classified by risk.
- **Independent Validation**: Models must be validated by a function independent of the model's development and use. Validation includes conceptual soundness review, ongoing monitoring, and outcomes analysis.
- **Model Governance**: Establish board and senior management oversight, clear roles and responsibilities, and policies for model development, implementation, and use.
- **Documentation**: Maintain complete documentation of model logic, data sources, assumptions, limitations, and validation results.

Application to Agentic Systems: For agentic systems deployed in banking (e.g., iterative credit underwriting agents that gather information across multiple cycles, adapt criteria based on discovered patterns, and escalate edge cases), SR 11-7 requires documentation of all six operational properties. Unlike traditional one-shot credit models that execute fixed logic, agentic systems must document iteration logic (when does the system gather more data?), adaptation mechanisms (how do criteria evolve?), and termination conditions (when does it escalate to humans?).

Securities: FINRA Supervision and Algorithmic Trading. FINRA Rule 3110 requires broker-dealers to establish supervisory systems reasonably designed to achieve compliance with applicable laws and regulations. For firms using algorithmic trading systems or AI-driven investment recommendations, this means:

- **Pre-Deployment Testing**: Validate algorithms in a controlled environment before production use.
- **Ongoing Monitoring**: Continuously monitor for erroneous or manipulative behavior.
- **Risk Controls**: Implement automated controls (e.g., price collars, volume limits) to prevent runaway algorithms.
- **Supervisory Review**: Designate supervisors responsible for algorithm oversight and establish escalation procedures.

Audit: PCAOB Standards on Audit Evidence and Sampling. The PCAOB has not issued AI-specific guidance, but existing auditing standards apply. AS 1105 (Audit Evidence) establishes that the auditor is responsible for all audit evidence, regardless of source.

Application to Agentic Systems: For agentic audit systems (e.g., agents that iteratively refine sampling strategies based on discovered anomalies, adapt risk assessments as they review documentation, and terminate when coverage objectives are met), PCAOB standards require:

- Documentation of iteration logic: How does the system refine its sampling or analysis strategy across cycles?
- Documentation of adaptation mechanisms: When and why does the system adjust risk assessments or expand sample sizes?
- Documentation of termination criteria: What triggers the system to conclude its work or escalate to human auditors?
- Understanding the tool's methodology and assumptions across all six GPA+IAT properties.
- Professional skepticism maintained throughout—auditors cannot delegate professional judgment to autonomous systems.

AS 2315 (Audit Sampling) requires auditors to design samples that provide a reasonable basis for conclusions. Agentic sampling systems must document how iteration and adaptation enhance (rather than compromise) statistical validity.

3.5 Layer 4: AI-Specific Regulation

AI-specific regulation is emerging rapidly. We focus on the most comprehensive framework and notable U.S. developments:

EU AI Act: Risk-Based Tiering. The EU AI Act, which entered into force in August 2024, establishes a risk-based regulatory framework (European Parliament and Council 2024):

- **Prohibited Practices**: AI systems that pose unacceptable risks (e.g., social scoring by governments, real-time biometric identification in public spaces except narrow law enforcement exceptions, manipulative or harmful systems) are banned.
- **High-Risk Systems**: AI systems used in employment, education, credit assessment, law enforcement, critical infrastructure, and biometric identification are classified as high-risk. These systems must satisfy stringent requirements:
  - **Risk Management** (Article 9): Establish and maintain a risk management system throughout the AI system's lifecycle.
  - **Data Governance** (Article 10): Training, validation, and testing datasets must be relevant, representative, and free from bias to the extent possible.
  - **Logging** (Article 12): Maintain automatic recording of events (logs) to enable traceability.
  - **Transparency** (Article 13): Provide clear instructions for use, including capabilities, limitations, and expected performance.
  - **Human Oversight** (Article 14): Design systems to enable effective oversight, including the ability to override or interrupt the system.
  - **Accuracy, Robustness, Cybersecurity** (Article 15): Achieve appropriate levels of accuracy and resilience against errors, faults, and cyberattacks.
  - **Conformity Assessment** (Article 43): High-risk systems must undergo third-party conformity assessment before market placement (for certain categories) or internal assessment (for others).
- **Limited-Risk and Minimal-Risk Systems**: Lower-risk systems face transparency obligations (e.g., chatbots must disclose they are AI) but not the full high-risk requirements.

Penalties for non-compliance are severe: up to €35 million or 7% of global annual turnover for prohibited practices; up to €15 million or 3% for high-risk system violations. Organizations operating in or serving EU markets must assess whether their agentic systems fall within high-risk categories and implement Article 9-15 requirements.

U.S. State and Local AI Laws. In the absence of comprehensive federal AI legislation, U.S. states and cities are enacting targeted rules:

- **Colorado AI Act (HB 24-1142)**: Effective January 1, 2026, Colorado's law prohibits algorithmic discrimination—deployment of high-risk AI systems that result in unlawful differential treatment or impact based on protected classifications (Colorado General Assembly 2024). Deployers must conduct impact assessments documenting the system's purpose, data sources, intended benefits, known limitations, and measures to mitigate discrimination. A rebuttable presumption of compliance applies if deployers complete a reasonable impact assessment in good faith.
- **New York City Local Law 144 (Automated Employment Decision Tools)**: Effective since July 2023, NYC requires employers using AI for hiring or promotion to conduct annual bias audits, publish summary results, and notify candidates that an automated tool is in use. Employers must also allow candidates to request alternative evaluation processes.
- **California Privacy Rights Act (CPRA) and Proposed AI Legislation**: California has enacted data protection laws that indirectly regulate AI (e.g., CPRA's provisions on automated decision-making) and is considering comprehensive AI legislation addressing high-risk uses.

These patchwork requirements mean organizations must track regulatory developments across jurisdictions and tailor governance to the most stringent applicable standard.

3.6 Layer 5: Voluntary Governance Frameworks

Voluntary frameworks provide structured approaches to AI governance. Organizations often adopt multiple frameworks to address different audiences and objectives:

NIST AI Risk Management Framework (AI RMF 1.0). Published in January 2023, the NIST AI RMF is a flexible, voluntary framework for managing AI risks (National Institute of Standards and Technology 2023). It organizes activities into four functions:

- **Govern**: Establish organizational structures, policies, and accountability for AI risk management.
- **Map**: Identify context, stakeholders, and potential impacts of AI systems.
- **Measure**: Assess and benchmark AI system performance, including trustworthiness characteristics (fairness, transparency, accountability, safety, privacy, security).
- **Manage**: Allocate resources, implement risk treatments, and monitor effectiveness.

NIST AI RMF emphasizes trustworthiness characteristics and provides flexibility for organizations of different sizes and sectors. It is widely referenced by federal agencies, state regulators, and private-sector organizations as a baseline governance framework.

ISO/IEC 42001:2023 (AI Management Systems). ISO/IEC 42001 is an international standard for AI management systems, providing a certifiable framework (International Organization for Standardization 2023). It establishes requirements for establishing, implementing, maintaining, and continually improving an AI management system. Annex A provides 40+ AI-specific controls organized by category (data management, model development, deployment, monitoring).

ISO/IEC 42001 is especially relevant for organizations:

- Operating in EU markets (the standard is recognized as supporting EU AI Act compliance).
- Seeking third-party certification to demonstrate governance maturity.
- Requiring international recognition (ISO standards are globally accepted).

Certification typically costs $50,000-$150,000 and requires 3-6 months, depending on organizational size and maturity.

COBIT (IT Governance Framework). COBIT, developed by ISACA, is a comprehensive IT governance framework widely used by enterprises. COBIT 2019 includes guidance on emerging technologies, including AI. Organizations with mature IT governance often extend COBIT to cover AI systems rather than creating parallel structures.

COBIT is best suited for organizations seeking to integrate AI governance into existing enterprise IT governance rather than treating AI as a standalone domain.

SOC 2 Type II (Vendor Assurance). SOC 2 (Service Organization Control) is an auditing framework for service providers, especially SaaS vendors. SOC 2 Type II reports assess controls over security, availability, processing integrity, confidentiality, and privacy over a period of time (typically 6-12 months).

For organizations procuring AI tools from vendors, a SOC 2 Type II report provides independent assurance that the vendor has implemented and operated controls effectively. Many enterprises require SOC 2 reports as a condition of vendor contracts.

Framework Selection Logic. Organizations often layer frameworks:

- **Start with NIST AI RMF** for flexible internal governance (free, widely recognized, no certification requirement).
- **Add ISO/IEC 42001** if seeking certification, operating in EU markets, or facing customer demands for third-party assurance.
- **Integrate with COBIT** if mature IT governance structures exist.
- **Require SOC 2** from third-party AI vendors to validate their controls.

No single framework addresses all requirements. Layering enables organizations to satisfy general governance needs (NIST), achieve certification (ISO), integrate with enterprise governance (COBIT), and validate vendor controls (SOC 2).

3.7 Seven Common Controls Across Frameworks

Despite structural differences, all governance frameworks converge on seven common controls:

1. **Risk Assessment and Management**: Identify, assess, prioritize, and mitigate AI-related risks throughout the system lifecycle. Risk assessment is the foundation for all subsequent governance activities.
2. **Human Oversight**: Implement oversight mechanisms proportionate to system autonomy and risk. Human-in-the-loop (pre-approval for high-stakes decisions), human-on-the-loop (monitoring with intervention capability), or human-in-command (strategic oversight with emergency stop authority).
3. **Audit Logging and Traceability**: Maintain tamper-evident logs that capture inputs, outputs, decisions, and human interventions. Logs must enable reconstruction of decisions for audit, investigation, and regulatory review.
4. **Explainability and Transparency**: Provide stakeholders—users, auditors, regulators—with understandable information about how the system operates, what factors influence decisions, and what limitations exist. Explainability techniques must be validated for faithfulness (reflects actual model logic), completeness (material factors included), and usefulness (enables informed decisions).
5. **Vendor Management**: Assess, monitor, and manage third-party AI vendors. Vendor due diligence, contract negotiation, ongoing monitoring, and escalation procedures are essential because vendor risks cascade into organizational liability.
6. **Incident Response and Remediation**: Detect, triage, contain, investigate, remediate, and learn from AI system failures. Incident response must be rapid (fairness violations and safety failures require immediate action) and systematic (root cause analysis, notification, continuous improvement).
7. **Documentation and Record-Keeping**: Maintain comprehensive documentation of system purpose, design, data sources, validation results, deployment decisions, monitoring outputs, and incidents. Documentation supports audits, regulatory inquiries, and continuous improvement.

While frameworks differ in emphasis and structure, these seven controls represent governance universals. Section 2 (presented earlier) established how to calibrate control intensity based on system properties. Section 4 operationalizes these calibrated controls through technical architecture and organizational processes.

4 Implementation: Building Governance Systems

Section 2 established principles for calibrating control intensity. This section operationalizes those principles: how to design and implement risk assessment, audit logging, explainability, human oversight, vendor management, performance monitoring, and incident response. We focus on actionable guidance—what practitioners and governance teams actually build—illustrated through examples from legal, financial, and audit domains.

4.1 Risk Assessment as Foundation

All governance begins with risk assessment. Before deploying an agentic system, organizations must systematically identify harm scenarios, assess their likelihood and impact, document mitigations, and define reassessment triggers.

Risk Assessment Methodology. Effective risk assessment addresses six categories of AI-related harms:

- **Bias and Fairness**: Does the system produce discriminatory outcomes? Are protected classes disproportionately harmed?
- **Accuracy and Reliability**: Does the system produce correct outputs? What is the error rate? What are the consequences of errors?
- **Security**: Can adversaries manipulate inputs (prompt injection), poison training data, or exfiltrate sensitive information?
- **Privacy**: Does the system access, process, or disclose personal or confidential information inappropriately?
- **Safety**: Can system failures cause physical harm, financial loss, or operational disruption?
- **Compliance**: Does deployment violate laws, regulations, or professional obligations?

For each risk category, assess *likelihood* (how probable is this harm?), *impact* (if it occurs, how severe are the consequences?), *affected stakeholders* (who is harmed?), and *mitigations* (what controls reduce risk?). Document *residual risk* after mitigations and obtain approval from appropriate governance authority (e.g., risk committee, general counsel, board for high-risk systems).

Define *reassessment triggers*: When must the risk assessment be updated? Common triggers include model updates, policy changes, regulatory developments, incident discoveries, and significant drift in performance or fairness metrics.

Example: Agentic Financial Planning Assistant Risk Assessment. A registered investment adviser deploys an agentic financial planning system that *iteratively* analyzes client portfolios, adapts recommendations based on market conditions and client feedback, and determines when to escalate to human advisers.

Agentic properties: The system has a goal (optimize client portfolio for risk-adjusted returns while satisfying regulatory constraints), perceives market data and client account state across multiple cycles, takes actions (generates rebalancing recommendations, requests additional client information), iterates through analysis-recommendation-feedback loops over days or weeks, adapts its strategy based on client responses and market changes, and terminates when confidence thresholds are met or escalation is required.

The risk assessment (condensed) identifies:

- **Compliance Risk** (High likelihood, High impact): System may recommend unsuitable investments violating Advisers Act fiduciary duty. Unlike a simple Q&A chatbot, this system *iterates* and *adapts*, creating compounding risk across multiple cycles. *Mitigation*: HITL approval required before any client-facing recommendation; compliance officer reviews recommendation logic monthly; system logs all intermediate reasoning steps for audit; quarterly fiduciary duty assessment by external counsel. *Residual risk*: Moderate.
- **Accuracy Risk** (Moderate likelihood, High impact): System may hallucinate market data or misinterpret client constraints across iterative cycles, leading to compounding errors. *Mitigation*: Ground all market data in verified sources (Bloomberg, Reuters APIs with cryptographic verification); implement cross-cycle consistency checks (flag contradictory recommendations); human adviser reviews final recommendation before client delivery; monthly accuracy testing against known portfolios. *Residual risk*: Moderate.
- **Adaptation Risk** (Moderate likelihood, High impact): System may adapt its strategy in ways that drift from regulatory compliance (e.g., learning to recommend higher-fee products based on firm incentives rather than client best interest). *Mitigation*: Adaptation limited to market analysis methods only; recommendation criteria remain fixed and auditable; quarterly adaptation audit reviews strategy changes; fee-based recommendations prohibited without explicit client authorization. *Residual risk*: Low-Moderate.
- **Iteration Risk** (Low likelihood, High impact): System may iterate excessively (analysis paralysis) or terminate prematurely (incomplete analysis). *Mitigation*: Define explicit termination conditions (maximum 5 iteration cycles OR confidence >0.85 OR 14-day timeout); log termination reason; human review if system terminates due to timeout rather than confidence. *Residual risk*: Low.
- **Security Risk** (Low likelihood, Very High impact): Prompt injection across iterative cycles could

manipulate accumulated state, causing disclosure of other clients' information. *Mitigation*: Input sanitization at each cycle; state integrity validation (cryptographic hashing); data isolation per client; monthly penetration testing focused on multi-cycle attacks. *Residual risk*: Low.

Monitoring: Daily review of active planning sessions (compliance), weekly adaptation log review, monthly accuracy and termination condition analysis, monthly security testing, continuous client feedback tracking.

This risk assessment demonstrates how agentic properties (iteration, adaptation, autonomous termination) create governance requirements beyond simple AI tools. The system's ability to iterate and adapt demands *cross-cycle consistency checks*, *adaptation audits*, and *termination condition validation*—controls unnecessary for non-agentic systems.

4.2 Audit Logging: Enabling Reconstruction and Accountability

Audit logging enables organizations to reconstruct decisions, investigate incidents, satisfy regulatory inquiries, and demonstrate accountability. Logging requirements scale with autonomy: high-autonomy systems (HIC) require more detailed logs than low-autonomy systems (HITL, where human review serves as primary control).

Logging Architecture Requirements. Effective audit logging captures:

- **Inputs**: What data did the system perceive? Include user queries, retrieved documents, API responses, sensor readings—whatever the system used to make decisions.
- **Outputs**: What did the system produce? Include recommendations, actions taken, messages sent, decisions rendered.
- **Decision Rationale**: Why did the system produce this output? For high-autonomy or high-consequence systems, log intermediate reasoning steps, confidence scores, alternative options considered.
- **Human Interventions**: When did humans approve, reject, or modify system outputs? Who made the decision? What was their rationale?
- **System State**: For stateful systems, log state changes to enable reconstruction of how the system's understanding evolved.

Logs must be stored in *tamper-evident* formats (e.g., append-only databases, cryptographic hashing) with access controls limiting who can read or delete logs. Retention periods must satisfy regulatory requirements: 7-10 years for financial services, 25 months minimum for ECOA adverse action records, potentially longer for litigation hold purposes.

Example: Agentic Credit Underwriting Audit Logging (ECOA Compliance). A bank deploys an agentic mortgage underwriting system that *iteratively* investigates applications by requesting additional documentation, querying third-party data sources (employment verification, asset verification), analyzing trends across multiple applicants, adapting its investigation strategy based on discovered risk patterns, and terminating when sufficient information is gathered or escalation is required. **Agentic properties**: Goal (approve qualified applicants while managing credit risk and satisfying ECOA requirements), Perception (observes application data, third-party verification responses, historical default patterns), Action (requests documents, queries APIs, generates preliminary assessments), Iteration (operates across 3-7 investigation cycles over 5-15 days), Adaptation (adjusts investigation depth based on risk indicators and application complexity), Termination (explicit conditions: confidence >0.90, maximum 7 cycles, or red-flag escalation to senior underwriter).

Equal Credit Opportunity Act Regulation B requires lenders to provide "principal reasons" for adverse credit decisions (Consumer Financial Protection Bureau 2011). For agentic systems that iterate across multiple cycles and adapt their investigation strategy, the logging architecture must capture *cross-cycle decision evolution* to enable reconstruction of how the system's assessment changed over time.

Listing: Agentic Underwriting Audit Log (Simplified JSON)

```
{
  "application_id": "APP-2024-00123",
  "session_start": "2024-11-20T14:32:15Z",
  "session_end": "2024-11-28T09:15:42Z",
  "model_version": "agentic-underwriting-v2.1",
  "total_cycles": 4,
  "termination_reason": "confidence_threshold_met",
  "cycles": [
    {
      "cycle": 1,
      "timestamp": "2024-11-20T14:32:15Z",
      "perception": ["application_form", "credit_report"],
      "action": "request_employment_verification",
      "preliminary_assessment": "UNCERTAIN",
      "confidence": 0.62,
    "rationale": "Initial DTI borderline; need employment stability confirmation"
    },
    {
      "cycle": 2,
      "timestamp": "2024-11-22T10:18:33Z",
      "perception": ["employment_verification_response"],
      "action": "request_asset_documentation",
      "preliminary_assessment": "UNCERTAIN",
```

```
      "confidence": 0.71,
    "rationale": "Employment stable; need asset verification for down payment"
    },
    {
      "cycle": 3,
      "timestamp": "2024-11-25T15:42:09Z",
      "perception": ["bank_statements", "investment_accounts"],
      "action": "analyze_comparable_approvals",
      "preliminary_assessment": "LIKELY_APPROVE",
      "confidence": 0.84,
     "rationale": "Assets verified; comparable risk profile to approved cases"
    },
    {
      "cycle": 4,
      "timestamp": "2024-11-28T09:15:42Z",
    "perception": ["market_conditions", "portfolio_concentration_analysis"],
      "action": "generate_final_recommendation",
      "final_decision": "APPROVE",
      "confidence": 0.92,
      "recommendation": "Approve with standard terms"
    }
  ],
  "final_decision_factors": [
    {"factor": "verified_employment_stability", "weight": 0.35},
    {"factor": "sufficient_liquid_assets", "weight": 0.30},
    {"factor": "comparable_risk_profile", "weight": 0.25},
    {"factor": "credit_score_within_guidelines", "weight": 0.10}
  ]
}
```

Retention: 25 months (ECOA requirement) + 7 years (standard banking litigation hold).

Security: Logs encrypted at rest; access restricted to compliance officers, auditors, and authorized investigators; append-only with cryptographic integrity verification (tamper-evident); per-cycle hash chains to detect any alteration.

Retrievability: Indexed by application ID, applicant (hashed identifier to protect PII), decision date, termination reason, and number of cycles. Enables compliance officers to query: “Show all adverse decisions where the system terminated due to timeout rather than confidence” or “Identify applications where preliminary assessment changed from LIKELY_APPROVE to ADVERSE between cycles.”

Validation: Quarterly audit sampling verifies logs enable reconstruction of iterative decision evolution; test whether system's cross-cycle adaptations comply with fair lending principles; validate termination conditions are consistently applied.

This logging architecture satisfies ECOA's explainability requirement while addressing agentic-specific concerns: it captures *how* the system's understanding evolved across cycles, *what* triggered adaptation, and *why* the system terminated. Without cross-cycle logging, the bank cannot reconstruct agentic decision-making or demonstrate that adaptation did not introduce prohibited discrimination.

4.3 Explainability: From Technical Outputs to Stakeholder Understanding

Explainability translates system behavior into understandable information for stakeholders—users, auditors, regulators, affected individuals. Regulatory requirements vary: ECOA requires "principal reasons," GDPR requires "meaningful information about the logic involved," PCAOB requires auditors to document the rationale for audit procedures. Explainability techniques must be selected based on regulatory requirements and validated for *faithfulness* (reflects actual model logic), *completeness* (material factors included), and *usefulness* (enables informed decisions).

Example: Agentic Audit Investigation System (PCAOB Compliance). A Big Four accounting firm develops an agentic audit assistant that *iteratively* investigates high-risk accounts receivable by analyzing transactions, requesting supporting documentation, cross-referencing with third-party data, adapting its investigation strategy based on discovered anomalies, and escalating to senior auditors when material issues are identified. **Agentic properties**: Goal (identify material misstatements in accounts receivable while satisfying PCAOB professional standards), Perception (observes transaction data, aging reports, payment history, correspondence, bank confirmations), Action (requests documents, flags anomalies, generates preliminary risk assessments, escalates findings), Iteration (conducts 2-5 investigation cycles per high-risk account over 1-3 weeks), Adaptation (adjusts investigation depth based on discovered red flags and audit evidence quality), Termination (explicit conditions: sufficient audit evidence obtained, material issue requiring escalation, or maximum cycle limit reached).

PCAOB Auditing Standards require auditors to design procedures that provide a reasonable basis for conclusions and to document the rationale in workpapers (Public Company Accounting Oversight Board 2010c; Public Company Accounting Oversight Board 2010d). For agentic systems that iteratively investigate and adapt their strategy, explainability must capture *why* the system escalated certain accounts, *how* its strategy evolved across cycles, and *what* evidence supported termination decisions.

System Design (Iterative Investigation with Explainable Adaptation):

1. **Cycle 1 (Initial Risk Scoring)**: The system applies documented risk criteria to identify high-risk receivables:
    - High-value (>$500K): Risk score +3

- Overdue >90 days: Risk score +2
- New customer (<1 year): Risk score +1
- Prior audit adjustments: Risk score +2
- Related-party transaction: Risk score +3

Accounts scoring ≥6 trigger iterative investigation.

2. **Cycle 2-N (Adaptive Investigation)**: For each high-risk account, the system:
   - Requests supporting documents (invoices, shipping confirmations, customer correspondence).
   - Analyzes payment patterns (unusual delays, partial payments, disputes).
   - Cross-references with third-party data (credit reports, public filings, industry payment norms).
   - *Adapts strategy*: If documentation is incomplete, requests additional evidence; if payment disputes detected, flags for legal review; if patterns suggest fraud, escalates immediately.
   - Generates cycle-level explanations: "Cycle 2: Requested shipping confirmations due to large value and customer dispute notation. Cycle 3: Shipping docs received but show delivery to alternate address—escalating for senior auditor review (potential revenue recognition issue)."
3. **Termination and Escalation**: The system terminates investigation when:
   - Sufficient audit evidence obtained (confidence >0.85).
   - Material issue identified requiring human escalation (red flag detected).
   - Maximum cycles reached (5 cycles) without resolution.

   Termination reason is logged and explained in workpapers.

Explainability Validation:

- **Faithfulness**: Verify explanations match actual investigation logic by reviewing audit logs (do logged perceptions and actions align with explanations?).
- **Completeness**: Confirm all material risk indicators that triggered escalation appear in explanations.
- **Usefulness**: Senior auditor reviews cycle-level explanations and confirms they enable professional judgment ("Does the system's escalation rationale justify senior auditor involvement?").

Workpaper Documentation: The audit workpaper includes:

- Initial risk scoring methodology (Cycle 1 criteria).

- Cycle-by-cycle investigation narrative (what the system perceived, what actions it took, why it adapted).
- Escalation rationale (why this account required human review).
- Senior auditor's assessment: "We deployed an agentic audit assistant to investigate 47 high-risk receivables. The system iteratively gathered evidence across 2-5 cycles per account, adapting its strategy based on discovered documentation quality and anomaly patterns. It escalated 8 accounts for senior review due to identified red flags (revenue recognition concerns, collectability doubts). We reviewed the system's investigation logs, assessed the escalated accounts, and obtained sufficient appropriate audit evidence to support our conclusions."

This agentic design satisfies PCAOB's requirement that auditors understand their methodology while demonstrating how iteration and adaptation improve audit effectiveness. The system's ability to *learn* during investigation (adapting strategy based on discovered evidence) and *escalate appropriately* (terminating when human judgment is required) exemplifies agentic governance in practice.

4.4 Human Oversight: Workflows for HITL, HOTL, and HIC

Section 2.2 defined three oversight modes. This section operationalizes them through workflows, notification mechanisms, intervention interfaces, and escalation procedures.

HITL (Human-in-the-Loop): Approval Workflows. HITL systems require human pre-approval before executing high-consequence actions. Implementation requires:

- **Approval Queue**: System generates a recommendation and adds it to a queue visible to authorized reviewers.
- **Notification**: Alert the reviewer (email, dashboard notification, SMS for time-sensitive actions).
- **Review Interface**: Present the recommendation, supporting evidence, system confidence, and options (approve, reject, modify, request more information).
- **Accountability**: Log who approved, when, and any modifications made.

Automation Bias Mitigation: To prevent rubber-stamping, randomize the presentation order of recommendations, periodically inject known-incorrect recommendations as controls, and track approval/rejection rates per reviewer (flag reviewers with suspiciously high approval rates).

HOTL (Human-on-the-Loop): Monitoring and Intervention. HOTL systems operate autonomously but humans monitor and can intervene. Implementation requires:

- **Monitoring Dashboard**: Real-time or near-real-time display of system activity (actions taken, error rates, escalation triggers, user feedback).

- **Escalation Triggers**: Define conditions requiring human review (e.g., low-confidence decisions <0.7, user complaints, outcomes near policy boundaries, anomalies detected).
- **Intervention Protocol**: How does the human halt the system, override a decision, or modify parameters? Must be accessible in real-time.
- **Escalation Pathway**: If the monitoring human cannot resolve an issue, to whom do they escalate? (Senior supervisor, compliance officer, emergency stop authority.)

Example: Agentic Credit Underwriting HOTL Monitoring. A mortgage lender's agentic underwriting system (described in Section 4.2) operates in HOTL mode, iteratively investigating applications across multiple cycles. Senior underwriters monitor aggregate system performance and intervene when agentic-specific escalation triggers fire:

- System confidence <0.70 after maximum cycles (7) → Escalate to senior underwriter for manual completion
- System terminates due to timeout rather than confidence → Human review of investigation adequacy
- Applicant disputes preliminary assessment → Human underwriter reviews cycle-by-cycle investigation log
- System's adaptation creates contradictory assessments across cycles → Flag for quality assurance review
- Fairness metric (monthly review) shows disparate impact >20% → Escalate to Chief Risk Officer; halt system pending investigation

Underwriters access a dashboard showing daily metrics specific to agentic operations: average cycles per application, termination reason distribution (confidence vs. timeout vs. red-flag escalation), adaptation frequency, cross-cycle consistency score, and fairness metrics.

If escalation frequency spikes or average cycles increase significantly, supervisors investigate root cause (data quality degradation, overly conservative termination thresholds, or emerging risk patterns requiring strategy adjustment).

HIC (Human-in-Command): Strategic Oversight and Emergency Stop. HIC systems operate with high autonomy. Humans set goals and constraints, monitor aggregate performance, and retain emergency stop authority. Implementation requires:

- **Strategic Goal-Setting**: Executives define objectives, risk appetite, and constraints (e.g., "Fraud detection system must achieve 95% precision, maintain false positive rate <1%, and satisfy GDPR Article 22 requirements").

- **Aggregate Monitoring**: Statistical dashboards (daily/weekly/monthly) showing performance trends, fairness metrics, error rates, drift indicators. Not individual-decision review.
- **Emergency Stop**: Accessible to authorized personnel (CTO, Chief Risk Officer, compliance head); tested quarterly; documented procedures for graceful shutdown (complete in-progress transactions, notify affected users, preserve state).
- **Revalidation Triggers**: Define when the system must be revalidated before continuing operation (e.g., fairness violation detected, accuracy below SLA, regulatory policy change).

4.5 Vendor Management: Assessing and Monitoring Third-Party AI

Most organizations procure AI systems from vendors rather than building in-house. Vendor risk cascades into organizational liability: if the vendor's model hallucinates, is biased, or breaches confidentiality, the deploying organization faces regulatory penalties and reputational harm. Governance must include vendor due diligence, contract negotiation, and ongoing monitoring.

Vendor Due Diligence Framework (Three Phases). **Phase 1: Initial Assessment (Questionnaire)**. Request documentation on:

- **Data Sources**: What training data was used? Is it proprietary, licensed, or scraped? How frequently is it updated? Does it include customer data from other clients (multi-tenancy risk)?
- **Model Architecture**: Is the model proprietary or open-source? What explainability techniques are available? Can the vendor provide confidence scores or uncertainty estimates?
- **Security and Confidentiality**: How is customer data handled? Is it used for training? Where is it stored (jurisdiction)? What encryption standards apply? SOC 2 certification?
- **Accuracy and Performance**: What benchmarks has the vendor tested? What is the error rate? In what domains/populations does performance degrade?
- **Bias and Fairness**: Has the vendor conducted fairness testing? What mitigation techniques are implemented? Can the vendor provide disaggregated performance metrics by protected class?

Phase 2: Document Review. Request and review:

- SOC 2 Type II report (if available)
- Data Processing Agreement (DPA) for GDPR compliance
- Model validation reports (accuracy, fairness, robustness)
- Security certifications (ISO 27001, FedRAMP, etc.)
- Sample explanations/outputs

Phase 3: Reference Checks and Pilot Testing. Contact existing clients in similar domains. Conduct pilot testing with representative data to validate accuracy, explainability, and performance claims.

Contract Negotiation: Shifting Risk to Vendors Where Possible. Negotiate contract terms that allocate risk appropriately:

- **Liability Caps**: Vendors typically propose caps (e.g., "Liability limited to fees paid in prior 12 months"). For high-risk use cases (credit decisioning, legal advice, audit), negotiate higher caps or uncapped liability for confidentiality breaches and gross negligence.
- **Model Update Notification**: Require advance notice (30-60 days) before material model updates, enabling the organization to revalidate before deployment.
- **Audit Rights**: Reserve the right to audit vendor controls annually or upon incident discovery.
- **Data Handling**: Prohibit use of customer data for training; require data deletion upon contract termination; specify jurisdiction for data storage.
- **SLAs**: Define performance thresholds (accuracy, uptime, response time); specify remedies for SLA violations.

Agentic-Specific Risk: Adaptation Opacity. Agentic systems that learn and adapt create a unique vendor risk that traditional AI contracts do not address: **adaptation opacity**—the vendor's model silently updates its decision-making strategy in the background without formal version changes, invalidating continuous validation requirements and creating regulatory exposure.

The Problem: Regulatory frameworks like SR 11-7 (Federal Reserve model risk management) require ongoing validation of models used by banking institutions (Board of Governors of the Federal Reserve System 2011). Organizations validate "Model v2.1" and deploy it. If the vendor's agentic system *adapts*—adjusting feature weights, refining decision criteria, or modifying iteration logic—the deployed system may behave materially differently from the validated version, yet the vendor does not issue a new version number or notify the customer. The organization continues operating under the assumption it is using validated "v2.1," but the system's actual behavior has drifted. This breaks continuous validation, exposes the organization to regulatory penalties ("You deployed an unvalidated model"), and creates fairness risk (adaptation may introduce prohibited discrimination).

Why Traditional Contracts Fail: Standard AI vendor contracts address *formal version updates* ("Vendor will notify Customer of material updates"). But agentic systems' adaptation mechanisms operate *within* a version, not across versions. The vendor's position: "We did not update the model—v2.1 is still v2.1. The system is designed to adapt; that is a feature, not a bug." The customer's regulatory obligation: "We must validate material changes to model behavior, regardless of version numbering."

Contractual Mitigation—Adaptation Transparency Clauses: For agentic vendor systems, negotiate specific contractual provisions that address adaptation opacity:

1. **Adaptation Mechanism Disclosure (Pre-Contract)**:
   - Vendor must disclose all adaptation mechanisms in the system (e.g., "Model adjusts feature weights based on prediction accuracy feedback; updates occur daily with exponential moving average decay factor 0.95").
   - Vendor must identify which components adapt (feature weights, decision thresholds, iteration limits, termination criteria, tool selection logic) and which remain static.
   - Vendor must provide technical documentation sufficient for independent validation of adaptation logic.
2. **Change Log Access (Operational)**:
   - Vendor must maintain detailed change logs tracking all adaptation events: what changed, when, by how much, and why (what feedback triggered the adaptation).
   - Customer must have API or dashboard access to change logs (daily or weekly exports).
   - Change logs must be retained for the longer of: contract term + 2 years, or applicable regulatory retention period (7-10 years for banking; 25 months for ECOA).
3. **Material Change Thresholds and Notification (Trigger-Based)**:
   - Define *material behavioral change* thresholds that trigger mandatory customer notification and revalidation rights. Recommended thresholds:
     - Feature weight adjustments: Any single feature weight changes by >10% (absolute) within a 30-day period.
     - Decision boundary shifts: Approval/rejection threshold changes by >5% within a 30-day period.
     - Performance degradation: Accuracy, precision, or recall degrades by >5% (absolute) on validation dataset.
     - Fairness drift: Disparate impact ratio changes by >10% for any protected class.
   - Upon threshold breach, vendor must: (a) notify customer within 5 business days; (b) provide root cause analysis; (c) offer rollback to prior configuration; (d) pause further adaptation pending customer approval to continue.
   - Customer retains right to require revalidation (at vendor's expense) or revert to last validated configuration.

4. **Audit and Testing Rights**:
    - Customer may conduct quarterly "behavioral validation" testing: submit test cases to the production system and compare outputs to baseline validated behavior.
    - If behavioral drift detected (outputs differ from validated baseline by >materiality threshold), customer may demand vendor investigation and remediation.
    - Vendor must cooperate with third-party audits of adaptation mechanisms (ISO 42001, SOC 2, or regulatory examination).
5. **Adaptation Freeze Option**:
    - Customer may request "adaptation freeze"—vendor disables learning mechanisms, and the system operates with static parameters.
    - Use case: During regulatory examination, high-risk deployment, or incident investigation, customer needs behavioral stability. Vendor must support freeze mode within 48 hours of request.

Example Contractual Language:

> **Section X: Adaptation Transparency and Change Control**
>
> **X.1 Adaptation Disclosure.** Vendor has disclosed in Exhibit C all mechanisms by which the System adapts its decision-making logic, including feature weight updates, threshold adjustments, and strategy refinements. Vendor represents that Exhibit C is complete and accurate as of the Effective Date.
>
> **X.2 Change Logs.** Vendor shall maintain detailed change logs documenting all adaptation events, including timestamp, changed parameters, magnitude of change, and triggering feedback. Customer shall have API access to change logs with daily refresh.
>
> **X.3 Material Change Notification.** If any of the following thresholds are met, Vendor shall notify Customer within five (5) business days and provide root cause analysis: (a) any feature weight changes by more than ten percent (10%) absolute within thirty (30) days; (b) decision threshold changes by more than five percent (5%) within thirty (30) days; (c) accuracy degrades by more than five percent (5%) on validation dataset; or (d) disparate impact ratio for any protected class changes by more than ten percent (10%).
>
> **X.4 Revalidation Rights.** Upon Material Change notification, Customer may elect to: (a) require Vendor to revert System to last validated configuration (at no cost to Customer); (b) conduct revalidation testing (Vendor shall cooperate and bear reasonable costs); or (c) pause System operation pending resolution.
>
> **X.5 Adaptation Freeze.** Upon forty-eight (48) hours' notice, Customer may require Vendor to disable all adaptation mechanisms, causing the System to operate with static parameters. Vendor shall maintain freeze mode for up to ninety (90) days per Calendar Year at no additional cost.

Governance Benefit: These contractual provisions operationalize continuous validation requirements for adaptive agentic systems. Without adaptation transparency, organizations deploying vendor agentic systems face a compliance gap: regulatory obligations demand ongoing validation, but vendor opacity prevents detection of material changes. Adaptation transparency clauses shift this burden back to vendors and provide customers with the visibility necessary to satisfy SR 11-7, ECOA, and similar frameworks.

Historical Parallel: Learning from Manufacturing Quality Evolution

The adaptation opacity challenge has a precedent: the evolution of quality control in manufacturing. Before the 1950s, quality control was *inspection-based*—manufacturers caught defects after production rather than preventing them. Small process changes (temperature adjustments, material substitutions) occurred without systematic tracking. Quality degraded invisibly until failures became visible to customers.

W. Edwards Deming introduced Statistical Process Control (SPC) to Japanese manufacturers in the 1950s: **measure process variation continuously**, not just inspect final products. This wasn't immediately adopted—it required new capabilities (statistical training, measurement infrastructure, documentation systems) that most organizations lacked. But manufacturers that built these capabilities gradually (Toyota, Sony) discovered quality improvement created competitive advantage. By the 1980s, SPC became standard practice, formalized in ISO 9000 (1987).

The AI opportunity: Adaptation transparency for agentic systems follows a similar trajectory. Most organizations today lack the capability to track behavioral drift systematically—just as most 1950s manufacturers lacked statistical quality control capabilities. The contractual clauses above represent *emerging best practice*, not current universal standard. Organizations can build this capability incrementally:

- **Start**: Require vendors to disclose adaptation mechanisms (even if full change log access is negotiated later).
- **Develop**: Pilot quarterly behavioral validation testing with one vendor system before scaling.
- **Mature**: Negotiate material change thresholds and adaptation freeze options as organizational validation capability grows.

Vendors developing adaptation transparency capabilities today—change logs, behavioral drift metrics, customer notification protocols—are positioned to demonstrate governance maturity as regulatory expectations crystallize. This is a **capability-building opportunity** for both vendors and deployers: just as SPC became a quality differentiator in manufacturing, adaptation transparency will become a governance differentiator in AI markets. Organizations that build these capabilities early will be prepared as standards emerge; those that delay will face retrofitting costs when regulatory or market pressure intensifies.

Ongoing Monitoring. Vendor due diligence does not end at contract signature. Implement:

- **Performance Monitoring**: Track accuracy, error rates, user complaints. Compare vendor claims to observed performance.
- **Security Monitoring**: Review vendor security incident reports; conduct annual security assessments.
- **Accuracy Audits**: Quarterly or semi-annual testing of vendor outputs against ground truth.
- **Escalation Procedures**: Define error rate thresholds triggering vendor review (e.g., "If hallucination rate exceeds 5%, escalate to General Counsel; consider vendor termination").

Example: Law Firm Foundation Model Vetting. A law firm evaluates a foundation model vendor for legal research assistance. Due diligence identifies five risk categories:

- **Confidentiality**: Vendor uses multi-tenant architecture; customer queries may be logged for training. *Mitigation*: Negotiate zero-retention DPA; require vendor to delete all firm data within 30 days of session termination; annual audit rights.
- **Conflicts**: Vendor serves competing law firms; could create conflicts if data is shared. *Mitigation*: Vendor affirms data isolation per client; third-party audit confirms isolation controls.
- **Accuracy**: Vendor claims 95% citation accuracy but provides no independent validation. *Mitigation*: Firm conducts pilot testing with 200 known cases; achieves 60% accuracy (below acceptable threshold). Vendor contract includes accuracy SLA (90%); quarterly accuracy audits; right to terminate if SLA violated for two consecutive quarters.
- **Hallucination**: Model occasionally fabricates case law. *Mitigation*: HITL verification (attorney must independently verify all citations before filing); firm maintains hallucination log; if hallucination rate >5%, escalate to General Counsel.
- **Regulatory Compliance**: ABA Rule 1.6 confidentiality obligations. *Mitigation*: Vendor contract includes uncapped liability for confidentiality breaches; cyber insurance confirmation.

Firm approves vendor with conditions: HITL verification mandatory, quarterly accuracy audits, annual security review, zero-retention DPA. This risk-calibrated approach enables use while protecting against residual vendor risks.

4.6 Performance Monitoring and Incident Response

Governance is not a one-time validation but a continuous cycle. Systems must be monitored for performance degradation, fairness violations, data drift, and security incidents. When failures occur, organizations must detect, contain, investigate, remediate, and learn.

Performance Monitoring: Four Dimensions. Monitor continuously across four dimensions:

1. **Performance Metrics**: Accuracy, precision, recall, F1 score, latency—whatever aligns with business objectives. Establish SLAs and alert when performance degrades below thresholds.
2. **Data Drift**: Are input distributions changing? If the system was trained on 2020-2022 mortgage applications and is now seeing 2024 applications with different characteristics (higher interest rates, different applicant demographics), performance may degrade.
3. **Concept Drift**: Are input-output relationships changing? For example, fraud patterns evolve; a fraud detection model trained on 2022 patterns may miss 2024 attack vectors.
4. **Fairness Metrics**: For systems affecting protected classes, monitor approval rates, error rates, and disparate impact ratios by demographic group. ECOA requires lenders to monitor for disparate impact; GDPR Article 22 requires ongoing assessment of automated decision-making.

Incident Response Cycle. When failures occur, follow a systematic cycle:

1. **Detect**: Monitoring alerts, user complaints, audit findings, or external reports identify a potential issue.
2. **Triage**: Assess severity. Is this a fairness violation (immediate board escalation and system halt)? Accuracy degradation (investigate root cause)? Isolated error (document and monitor)? Security breach (activate incident response team)?
3. **Contain**: Limit harm. For fairness violations or safety failures, halt the system immediately. For accuracy degradation, consider reverting to prior known-good model version. For security breaches, isolate affected systems.
4. **Investigate**: Root cause analysis. What caused the failure? Data quality issue? Model drift? Adversarial attack? Process breakdown (human failed to review)?
5. **Remediate**: Fix the root cause. Retrain model, update data sources, patch vulnerability, revise process.
6. **Notify**: Regulatory notification (if required), affected individuals (if harm occurred), internal stakeholders (board, executives).
7. **Post-Incident Review**: Document lessons learned, update risk assessment, revise controls to prevent recurrence.

Example: Disparate Impact Detected in Agentic Credit Underwriting. This example demonstrates the incident response cycle applied to an agentic-specific fairness failure:

1. **Detect**: A bank's monthly fairness review identifies disparate impact in its agentic mortgage underwriting system. Hispanic applicants have a 65% approval rate compared to 82% for white applicants (80% rule violated: 65/82 = 79.3%). The compliance monitoring dashboard flags this violation automatically and escalates to the Chief Risk Officer.
2. **Triage**: The Chief Risk Officer assesses severity. This is a fairness violation under ECOA—a high-severity incident requiring immediate action. ECOA violations create strict liability (intent not required), expose the bank to regulatory penalties (CFPB enforcement action), civil litigation (class action risk), and reputational harm. The CRO escalates to the board and activates the fairness incident response protocol.
3. **Contain**: The bank immediately halts the agentic underwriting system and reverts to manual underwriting pending investigation. All pending applications in the system are reassigned to human underwriters. The compliance team preserves all cycle-level logs for the past 12 months for forensic analysis (read-only archive to prevent tampering). No new applications enter the agentic system until root cause is identified and remediated.
4. **Investigate**: Because this is an agentic system that iterates and adapts, the investigation examines *cross-cycle bias accumulation*—whether discrimination emerges through the system's iterative investigation strategy rather than a single scoring function. Cycle-by-cycle log analysis reveals:
   - The system does not use discriminatory features (race, national origin) in initial risk assessment (Cycle 1). Cycle 1 scoring is facially neutral and satisfies fairness testing.
   - However, the system's *adaptive investigation strategy* creates disparate impact across subsequent cycles:
     - Hispanic applicants trigger more aggressive verification cycles (average 5.2 cycles vs. 3.8 for white applicants).
     - The system adapted to request employment verification more frequently for applicants with shorter U.S. employment tenure—a proxy for national origin, prohibited under ECOA.
     - Prolonged investigation cycles correlate with application abandonment: applicants withdraw during extended verification (28% abandonment rate for Hispanic applicants vs. 12% for white applicants), depressing approval rates even when the system would ultimately approve.
   - Root cause: The system's adaptation logic learned that shorter U.S. employment tenure correlated with higher default risk in historical training data. It adapted to investigate these applicants more aggressively. While statistically predictive, this created *process-based*

discrimination—disparate treatment through investigation burden, not final decision criteria.

5. **Remediate**: The bank implements four remediation measures before redeploying the system:

    - **Adaptation constraints**: Prohibit the system from using employment tenure (U.S. or total) as a factor in determining investigation depth. Constrain adaptation to prevent proxies for national origin from influencing cycle counts.
    - **Process parity monitoring**: Implement cycle-count parity monitoring across protected classes. Flag applications where investigation cycles deviate >20% from demographic-group median; escalate to human underwriter.
    - **Abandonment risk tracking**: Monitor application abandonment rates by protected class. Alert if abandonment rate for any protected class exceeds overall rate by >10 percentage points.
    - **Retrain and revalidate**: Retrain the adaptation logic with fairness constraints; validate fairness across *both* final outcomes (approval rates) and investigation process (cycle counts, abandonment rates) before redeployment.

6. **Notify**: The bank self-reports the ECOA violation to the Consumer Financial Protection Bureau (CFPB) within 30 days, providing root cause analysis and remediation plan. The bank notifies all Hispanic applicants processed during the affected period (past 12 months) and offers remediation: expedited re-review using the corrected system with mandatory human oversight, waived application fees for re-review, and priority processing.

7. **Post-Incident Review**: The compliance team conducts a structured post-incident review and implements systemic improvements:

    - **Risk assessment update**: Recognize *iteration bias* and *adaptation fairness* as distinct risk categories. Traditional fairness testing (outcome parity) is insufficient for agentic systems; process parity must be monitored separately.
    - **Monitoring enhancement**: Implement monthly cross-cycle fairness monitoring (not just final decision fairness). Dashboard now tracks average cycles by protected class, termination reason distribution, and abandonment rates.
    - **Validation protocol revision**: Add investigation process audits to pre-deployment validation. Future model updates must validate fairness across *both* outcomes and iteration behavior before production deployment.
    - **Training update**: Train underwriters and compliance officers on process-based discrimination risk specific to agentic systems.

This incident demonstrates a governance challenge unique to agentic systems: discrimination can emerge through *how* the system iterates and adapts, not just *what* it decides. Traditional fairness testing focuses on outcome parity; agentic governance must also ensure *process parity*.

---

5 Accountability and Organizational Structure

Technical controls alone do not create accountability. Governance requires explicit assignment of roles and responsibilities: who approves deployments, who monitors performance, who investigates incidents, who escalates to regulators? This section presents three organizational governance models, demonstrates role assignment through RACI matrices, defines escalation and reporting structures, and examines liability allocation. The goal is to ensure every governance activity has a clearly accountable owner.

5.1 Three Organizational Governance Models

Organizations structure AI governance in three primary ways, each with advantages and disadvantages depending on size, AI maturity, and regulatory intensity.

Centralized Model: Single AI Governance Office. A dedicated AI governance office or committee reports to senior leadership (typically the Chief Risk Officer, Chief Compliance Officer, or Chief Technology Officer), establishing policies, reviewing all proposed AI deployments, conducting risk assessments, and monitoring compliance. This model suits small to medium organizations (500-2,000 employees) with limited AI systems (5-20 use cases), high regulatory stakes (financial services, healthcare, legal), or early AI maturity where governance capability is being built.

Advantages:

- Consistency: Single office ensures uniform governance standards across all systems.
- Expertise concentration: Governance specialists develop deep knowledge of regulatory requirements and best practices.
- Clear accountability: One office owns all AI governance decisions.
- Easier audit: Regulators and internal auditors interact with a single governance function.

Disadvantages:

- Bottleneck risk: All deployment decisions route through one office, creating delays.
- Limited domain expertise: Central office may lack deep knowledge of domain-specific require-

ments (e.g., PCAOB audit standards, ECOA fair lending nuances).

- Scalability: As AI adoption grows, central office becomes overwhelmed.

Example: Regional investment advisory firm (500 employees, 10 AI tools) establishes AI Governance Office under Chief Compliance Officer with governance lead, technical specialist, and support staff conducting quarterly system reviews.

Federated Model: Central Coordination with Distributed Expertise. A central AI governance function establishes enterprise-wide policies and standards, while domain-specific governance teams (e.g., audit practice AI lead, tax practice AI lead, wealth management AI lead) implement and monitor compliance within their areas. The central function coordinates, audits federated teams, and escalates enterprise-wide issues. This model suits large organizations (5,000+ employees) with diverse AI use cases across multiple domains (50+ systems), mature AI adoption, and domain-specific regulatory requirements (audit, legal, banking, securities).

Advantages:

- Domain expertise: Practice leads understand PCAOB standards, tax regulations, or wealth management suitability rules better than a central office.
- Scalability: Distributed teams prevent central bottlenecks.
- Tailored governance: Each domain calibrates controls to specific regulatory and risk contexts.

Disadvantages:

- Inconsistency risk: Different domains may interpret policies differently or adopt varying standards.
- Coordination overhead: Central function must monitor multiple federated teams.
- Accountability diffusion: Harder to pinpoint responsibility when governance is distributed.

Example: Big Four accounting firm (10,000 employees, 50+ AI tools) establishes central AI Governance Committee setting firm-wide policies while each practice (audit, tax, advisory) designates domain-specific AI Leads ensuring compliance with practice-specific regulations (PCAOB, IRS, client confidentiality).

Embedded Model: Governance Within Existing Functions. AI governance is integrated into existing risk management, compliance, IT governance, and legal functions rather than creating a separate AI-specific structure, with each function applying its governance processes to AI systems. This model suits organizations with mature, well-functioning governance (strong ERM, compliance, IT governance), AI systems that extend existing processes (e.g., AI-enhanced fraud detection within existing fraud team), and leadership that prefers integration over new silos.

Advantages:

- Efficiency: Leverages existing governance infrastructure.
- Avoids silos: Prevents AI governance from operating in isolation from enterprise risk management.
- Cultural fit: Organizations resistant to new bureaucracy prefer extending existing processes.

Disadvantages:

- Expertise gaps: Existing functions may lack AI-specific knowledge (fairness testing, model validation, adversarial robustness).
- Accountability ambiguity: If AI governance is "everyone's responsibility," it may become no one's priority.
- Inconsistent application: Different functions may apply AI governance unevenly.

This model requires AI-specific training for existing governance personnel and clear assignment of AI oversight responsibilities within each function.

5.2 RACI Matrix: Operationalizing Accountability

Regardless of governance model, organizations must assign accountability for each governance activity using a RACI framework:

- **R (Responsible)**: Who does the work? (May be multiple people.)
- **A (Accountable)**: Who has decision authority and ultimate accountability? (*Only one A per activity.*)
- **C (Consulted)**: Who provides input or expertise before decisions?
- **I (Informed)**: Who is notified after decisions?

The key principle: **every governance activity must have exactly one Accountable party**. Diffused accountability ("the team is accountable") creates gaps where no one takes ownership.

Table 5 provides a sample RACI matrix for AI governance activities.

Key Observations from the Matrix. :

- **Single Accountability**: Each activity has one A. For example, the CRO (Chief Risk Officer) is accountable for fairness violation investigations; the AI Governance Lead is accountable for

Table 5: Sample RACI Matrix for AI Governance Activities

Activity	Board / CEO	CRO / CCO	AI Governance Lead	System Owner	Legal / Compliance
Approve enterprise AI governance policy	A	C	R	I	C
Approve low-risk AI deployment	I	I	A	R	C
Approve high-risk AI deployment	A	C	R	R	C
Conduct pre-deployment risk assessment	I	C	A	R	C
Monitor system performance (ongoing)	I	I	C	A, R	I
Investigate fairness violation	I	A	C	R	C
Approve vendor contract (high-risk system)	I	A	C	R	C
Report to board (quarterly AI governance update)	I	A	R	I	C
Respond to regulatory inquiry	C	A	R	R	A (legal)

low-risk deployments.

- **Escalation**: High-risk deployments elevate accountability to the Board/CEO, while low-risk deployments can be approved by the AI Governance Lead. This prevents bottlenecks (Board does not review every chatbot deployment) while ensuring senior oversight for consequential systems.
- **Multiple Responsible Parties**: Risk assessments may involve both the AI Governance Lead (methodological expertise) and the System Owner (domain knowledge). Both contribute, but only one is Accountable for the final approval.
- **Consultation and Information Flow**: Legal and Compliance are Consulted on most activities, ensuring regulatory considerations inform decisions. The Board is Informed of governance activities but not burdened with operational details.

Organizations should customize this matrix to their structure, size, and regulatory context. The principle—single accountability per activity—remains universal.

5.3 Escalation and Reporting

Governance requires clear escalation triggers: when must an operational issue be escalated to management, executives, or the board? And what cadence and format should governance reporting follow?

Three-Tier Escalation Model. **Tier 1 (Operational)**: Routine issues managed by system owners and AI governance teams without escalation. Examples: Low-confidence prediction (handled via HOTL intervention), minor accuracy fluctuation within SLA, user feedback complaint (resolved through customer service).

Tier 2 (Significant): Issues requiring management review and potential policy or system changes. Examples: Accuracy degradation below SLA for two consecutive weeks, vendor contract violation, user complaint alleging discrimination, data drift requiring model retraining. *Escalate to*: Chief Risk Officer, Chief Compliance Officer, or AI Governance Committee. *Response timeline*: Within 48-72 hours; investigate root cause and implement corrective action.

Tier 3 (Critical): Issues posing material risk to the organization, requiring immediate executive or board action. Examples: Fairness violation (disparate impact detected in credit decisioning), security breach (confidential data exfiltration), regulatory inquiry, systemic failure affecting multiple systems or customers. *Escalate to*: CEO, Board Risk Committee, General Counsel. *Response timeline*: Immediate (within hours); may require emergency board meeting, system halt, regulatory self-report.

Reporting Cadence and Audience. **Operational Dashboards (Daily/Weekly)**: System owners and AI governance teams monitor real-time or near-real-time dashboards showing performance metrics, error rates, escalation counts, user feedback. These are working tools, not executive reports.

Management Reports (Monthly/Quarterly): Chief Risk Officer and Chief Compliance Officer receive summary reports: number of systems deployed, risk assessments completed, incidents investigated, SLA compliance, vendor performance, upcoming regulatory developments. Format: 2-5 page executive summary with supporting appendices.

Board Presentations (Quarterly/Annual): Board receives narrative synthesis: strategic governance posture (are we ahead of or behind regulatory curve?), high-risk system approvals, material incidents and responses, policy changes, budget and resource requests. Format: 10-15 slide deck; focus on risk appetite alignment, not operational details.

Example Escalation: Fairness Violation in Credit Decisioning. A bank's monthly fairness monitoring detects disparate impact in credit pre-screening (see Section 4.6).

Tier 1 → Tier 3 Escalation:

- **Initial detection**: Compliance analyst (Tier 1 monitoring) identifies 80% rule violation.
- **Immediate escalation to Tier 3**: Fairness violations are pre-defined as critical (ECOA strict liability; regulatory penalties; class action risk). Analyst notifies Chief Risk Officer within 2 hours.
- **CRO response**: Halt system immediately (emergency stop protocol); convene incident response team (Legal, Compliance, AI Governance Lead, System Owner); notify CEO and Board Risk Committee within 24 hours.
- **Board involvement**: Emergency board meeting (if material); approve remediation plan (model retraining, affected applicant notification, CFPB self-report); allocate investigation budget.
- **Regulatory report**: General Counsel files self-report to CFPB within regulatory timeline.

This escalation pathway ensures the organization responds rapidly to critical risks and maintains board-level visibility into material governance failures.

5.4 Liability Allocation: Who Bears the Risk?

A foundational reality shapes AI governance: **liability concentrates on deployers, not vendors or technology**. Understanding this allocation is essential for calibrating governance investments.

Deployers Bear Primary Liability. When an AI system causes harm—discriminates against a protected class, provides inaccurate advice, breaches confidentiality—the deploying organization faces legal consequences:

- **Regulatory penalties**: ECOA violations, GDPR breaches, professional responsibility sanctions.
- **Civil liability**: Class actions, individual lawsuits, breach of fiduciary duty claims.

- **Reputational harm**: Client defection, loss of trust, negative publicity.

The fact that the system was purchased from a reputable vendor, relies on cutting-edge technology, or was approved by experts does not shield the deployer from liability. Professional duties (attorney competence, fiduciary obligations, auditor independence) are non-delegable.

Vendor Liability is Limited by Contract. Vendor contracts typically shift risk to deployers through:

- **Liability caps**: "Vendor's total liability shall not exceed fees paid in the prior 12 months." For a $50,000/year SaaS subscription, this caps vendor exposure at $50,000—insufficient to cover a $5 million ECOA class action settlement or $10 million GDPR penalty.
- **Warranty disclaimers**: "Vendor makes no warranties regarding accuracy, completeness, or fitness for a particular purpose." Deployers cannot recover damages for model hallucinations or bias if the vendor disclaimed such warranties.
- **Indemnification limits**: Vendors may indemnify only for certain risks (e.g., IP infringement) but exclude liability for "deployer's use of the system."

Governance as Primary Defense. Since deployers bear most liability and cannot fully recover from vendors, *governance becomes the primary defense*:

- **Regulatory defense**: Demonstrating reasonable care through documented risk assessments, monitoring, and incident response may reduce penalties or satisfy regulatory expectations.
- **Litigation defense**: Evidence of good-faith governance efforts may reduce damages, support summary judgment motions, or enable favorable settlements.
- **Insurance**: Insurers may require evidence of governance (policies, audits, controls) as a condition of coverage or premium reduction.

Organizations that deploy AI systems without governance face *uninsurable, unmitigated risk*. Conversely, robust governance creates an evidentiary record of due diligence—valuable in regulatory inquiries, litigation, and board oversight.

Example: Credit Decisioning Liability Chain. A mortgage applicant is denied credit by a bank using an AI underwriting system. The applicant sues under ECOA, alleging disparate impact (the system disproportionately denies applications from Hispanic applicants). The liability chain unfolds:

1. **Applicant sues bank**: ECOA imposes strict liability on the *creditor* (the bank), not the technology vendor. The bank is the defendant, regardless of whether it built the system in-house or purchased it.

2. **Bank investigates vendor recovery**: The bank's contract with the AI vendor caps liability at $100,000 (annual subscription fee). The ECOA settlement is $3 million (class action covering 500 affected applicants). The bank recovers only $100,000—3% of total damages.
3. **Bank disciplines employee**: The bank's AI governance policy required quarterly fairness monitoring. The assigned compliance analyst failed to conduct monitoring for six months. The bank terminates the analyst but remains liable to applicants and regulators (the analyst's failure does not excuse the bank's ECOA violation).
4. **Regulatory escalation**: The Consumer Financial Protection Bureau (CFPB) investigates and imposes a $5 million penalty for systemic ECOA violations. The penalty is assessed against the bank, not the vendor or employee.

Outcome: The bank bears $8 million in total liability ($3M settlement + $5M penalty) and recovers $100K from the vendor. Effective governance—quarterly fairness monitoring, documented risk assessment, incident response protocols—might have detected the bias earlier, limited exposure, and demonstrated good faith to regulators.

Liability Reality Check

"The AI did it" is not a legal defense. "We bought it from a reputable vendor" does not transfer liability. "Our employee was supposed to monitor it" does not excuse organizational failures. Deployers own the risk. Governance is the mechanism for managing it.

6 Examples in Context

This section demonstrates governance principles through worked examples in legal, financial, and accounting contexts. Each example follows a common governance framework: identify risks, calibrate controls, implement monitoring, and respond to incidents. These examples are illustrative—organizations must tailor governance to their specific regulatory obligations, risk appetite, and operational context—but they demonstrate how the conceptual frameworks from Sections 2 through 5 translate into practice.

6.1 Legal Domain: Professional Responsibility and Incident Management

Example 1: Agentic Legal Research Assistant—Iteration and Verification Controls. A mid-sized law firm deploys an agentic legal research system that *iteratively* investigates legal questions by formulating search strategies, retrieving cases, analyzing precedential value, cross-referencing citations, adapting its search based on relevance patterns, and terminating when sufficient authority is identified or confidence thresholds require human escalation. **Agentic properties**: Goal (identify controlling and persuasive authority), Perception (observes case law databases, statutes, treatises), Action (executes queries, extracts holdings, generates summaries), Iteration (conducts 2-6 research cycles, refining queries), Adaptation (adjusts search when initial results unpersuasive or contradictory authority discovered), Termination (confidence >0.85, maximum 6 cycles, or contradictory binding precedent requiring escalation).

Dimensional profile: HITL + human frame + static goals + stateless.

Incident: Cross-Cycle Hallucination Detected (Month 3): Opposing counsel alerts the firm that a motion contains citations that, while identifying real cases, mischaracterize holdings. Investigation reveals the system adapted its case selection across cycles in ways that introduced errors: early cycles identified correct cases, but later cycles synthesized holdings incorrectly when attempting to resolve contradictions.

Immediate Response (Detection and Containment):

- **Halt tool usage**: Firm suspends system access pending investigation.
- **Court notification**: Attorneys file corrected motion; submit candor-to-tribunal explanation (ABA Rule 3.3).
- **Client notification**: Firm notifies client; offers fee reduction.

Investigation (Agentic-Specific Root Cause):

- **Cycle-level audit**: Iteration logs reveal: Cycles 1-3 correctly identified 8 cases; Cycle 4 detected contradiction and attempted to "harmonize" holdings through paraphrasing—introducing mischaracterization; Cycles 5-6 propagated erroneous synthesis.
- **Adaptation failure**: System's contradiction-resolution logic created hallucination risk when paraphrasing rather than quoting verbatim.
- **Termination failure**: System terminated based on confidence >0.85 despite holding mischaracterization; confidence metric did not capture citation accuracy.
- **Sampling prior work**: 47 research sessions reviewed; 6 additional sessions where cross-cycle adaptation introduced errors (13% error rate).

Remediation (Agentic-Specific Controls):

- **Adaptation constraints**: Prohibit paraphrasing holdings; require verbatim quotations with Bluebook pin cites.

- **Cross-cycle consistency checks**: Flag when later cycles contradict earlier findings; escalate contradictions to attorney rather than automated resolution.
- **Termination condition revision**: Confidence must include citation verification score (database cross-check); terminate only if legal relevance confidence >0.85 AND citation accuracy >0.95.
- **HITL verification strengthened**: Attorneys must review cycle-by-cycle logs, not just final output; workpapers document validation of each cited case.
- **Quarterly iteration audits**: Sample 15% of research sessions; review cross-cycle adaptation for citation accuracy.

Governance Principles Demonstrated:

- **Agentic risk**: Iteration and adaptation compound errors across cycles; single-point checks insufficient.
- **Cross-cycle accountability**: Governance must audit how the system evolved, not just final outputs.
- **Termination calibration**: Confidence thresholds must account for domain-specific accuracy (citation fidelity).
- **Professional duty non-delegation**: Attorneys must understand iterative logic to fulfill Rule 1.1 competence obligations.

6.2 Finance Domain: Fair Lending and Fiduciary Duty

Example 2: Agentic Credit Underwriting—Process-Based Discrimination. A regional bank deploys an agentic mortgage underwriting system (detailed in Section 4.2) that *iteratively* investigates applications across 3-7 cycles over 5-15 days, adapting its investigation strategy based on discovered risk patterns. **Agentic properties**: Goal (approve qualified applicants while managing credit risk and satisfying ECOA), Perception (observes application data, third-party verification responses, historical default patterns), Action (requests documents, queries APIs, generates assessments), Iteration (multiple investigation cycles), Adaptation (adjusts investigation depth based on risk indicators), Termination (confidence >0.90, maximum 7 cycles, or red-flag escalation).

Dimensional profile: HIC + institutional frame + adaptive goals + stateful.

Agentic-Specific ECOA Compliance Challenges:

- **Cross-cycle logging**: Regulation B requires "principal reasons" for adverse decisions (Consumer Financial Protection Bureau 2011). For agentic systems, logs must capture how assessment evolved across cycles, not just final decision.
- **Process parity**: Beyond outcome fairness (80% rule), agentic systems require *process fair-*

ness—investigation cycles must not disproportionately burden protected classes.

- **Adaptation monitoring**: System's strategy adjustments must not introduce prohibited discrimination (e.g., learning to request more documentation from certain demographic groups).

Incident: Process-Based Discrimination Detected (Month 6): Monthly fairness review identifies Hispanic applicants have 65% approval rate vs. 82% for white applicants (79.3% ratio—violates 80% rule). Critically, cycle-by-cycle analysis reveals the system's *iterative investigation* creates disparate impact: Hispanic applicants trigger 5.2 average cycles vs. 3.8 for white applicants. The system adapted to request employment verification more frequently for shorter U.S. tenure (proxy for national origin—prohibited under ECOA). Prolonged cycles correlate with application abandonment. See Section 4.6 for complete incident response.

Governance Principles Demonstrated:

- **Agentic fairness risk**: Discrimination emerges through *how* the system iterates (process), not just final outcomes.
- **Cross-cycle accountability**: Traditional fairness testing (outcome parity) insufficient; must audit investigation process across cycles.
- **Adaptation constraints**: System's learning must be constrained to prevent adaptation from introducing prohibited proxies.
- **Termination parity**: Cycle-count monitoring ensures investigation burdens are distributed fairly across demographic groups.

Example 3: Agentic Financial Planning Assistant—Adaptation and Fiduciary Risk. A registered investment adviser deploys an agentic financial planning system (detailed in Section 4.1) that *iteratively* analyzes client portfolios, adapts recommendations based on market conditions and client feedback, and determines when to escalate to human advisers. **Agentic properties**: Goal (optimize portfolio for risk-adjusted returns while satisfying regulatory constraints), Perception (observes market data and client account state across multiple cycles), Action (generates rebalancing recommendations, requests additional client information), Iteration (analysis-recommendation-feedback loops over days or weeks), Adaptation (adjusts strategy based on client responses and market changes), Termination (confidence thresholds met or escalation required).

Dimensional profile: HITL + hybrid frame + adaptive goals + stateful.

Risk-Calibrated Controls (Agentic-Specific):

- **Compliance risk (High)**: Unlike simple Q&A chatbots, iterative adaptation creates compounding fiduciary risk. *Controls*: HITL approval before client-facing recommendation; compliance officer reviews recommendation logic monthly; system logs all intermediate reasoning for audit;

quarterly fiduciary duty assessment.

- **Adaptation risk (Moderate-High)**: System may adapt toward firm incentives (higher-fee products) rather than client best interest. *Controls*: Adaptation limited to market analysis methods only; recommendation criteria remain fixed and auditable; fee-based recommendations prohibited without explicit client authorization.
- **Iteration risk (Low-Moderate)**: System may iterate excessively (analysis paralysis) or terminate prematurely. *Controls*: Explicit termination conditions (maximum 5 cycles OR confidence >0.85 OR 14-day timeout); human review if timeout termination.
- **Cross-cycle consistency risk (Moderate)**: System may generate contradictory recommendations across cycles. *Controls*: Cross-cycle consistency checks; flag contradictory recommendations for human review.

Governance: Daily active session review (compliance), weekly adaptation log review, monthly termination condition analysis, continuous client feedback. Demonstrates how agentic properties (iteration, adaptation, termination) demand controls unnecessary for non-agentic tools.

6.3 Accounting Domain: Independence and Professional Skepticism

Example 4: AI Acceptable Use Policy for Agentic Systems (AICPA Independence). A Big Four accounting firm establishes an AI acceptable use policy to operationalize AICPA independence rules and SEC auditor independence requirements for *agentic audit and advisory systems*.

Dimensional profile: Spans HITL, HOTL, and HIC modes across human and institutional frames; policy-level governance rather than a single system.

Policy Structure (Condensed):

Scope: Applies to all AI tools used in audit, tax, and advisory engagements.

Guiding Principles:

- Independence: AI tools must not impair auditor independence (no management decision-making, no self-review threats).
- Competence: Professionals must understand tool capabilities and limitations (AICPA due care standard).
- Confidentiality: Client data must be protected; tools must satisfy data processing agreements.

Permitted Uses:

- Research and analysis (e.g., industry benchmarking, accounting standards research).
- Data analytics (e.g., anomaly detection, sampling optimization).

- Documentation assistance (e.g., workpaper summaries, drafted under professional review).

Prohibited Uses:

- Management decisions (e.g., AI selecting accounting policies for audit client—creates self-review threat).
- Audit opinions (e.g., AI drafting audit opinion without auditor's professional judgment application).
- Confidential data training (e.g., using client data to train vendor models without informed consent and DPA).

Safeguards:

- **Vendor requirements**: SOC 2 Type II report, Data Processing Agreement (GDPR-compliant), encryption (data at rest and in transit), PII minimization (use client codes, not full names).
- **Professional review**: Partners and managers must review AI-assisted work product before delivery to clients or inclusion in audit workpapers.
- **Documentation**: Workpapers must identify which procedures used AI tools, describe tool methodology, and explain professional judgment applied.

Training:

- **Role-based**: Auditors receive 4-hour training on tool capabilities, independence considerations, and documentation requirements. Partners receive 2-hour executive briefing.
- **Annual refresher**: 1-hour update on new tools, regulatory developments, lessons learned from incidents.

Incident Reporting:

- **Immediate escalation triggers**: Independence impairment suspected, confidential data breach, client complaint about AI tool, PCAOB or SEC inquiry.
- **Escalation pathway**: Report to Engagement Partner → National Office Ethics/Independence Group → General Counsel (if regulatory implications).

Governance Principles Demonstrated:

- **Domain-specific calibration**: Policy tailored to AICPA and SEC independence rules, not generic AI governance.
- **Role-based permissions**: Distinguishes permitted (research, analytics) from prohibited (man-

agement decisions, audit opinions) uses.

- **Accountability assignment**: Partners responsible for reviewing AI-assisted work; National Office Ethics Group accountable for policy updates.

Example 5: Agentic Audit Investigation System—Iterative Evidence Gathering. A Big Four firm deploys an agentic audit assistant (detailed in Section 4.3) that *iteratively* investigates high-risk accounts receivable, adapting its investigation strategy based on discovered anomalies and escalating to senior auditors when material issues are identified. **Agentic properties**: Goal (identify material misstatements while satisfying PCAOB standards), Perception (observes transaction data, aging reports, payment history, bank confirmations), Action (requests documents, flags anomalies, generates risk assessments), Iteration (conducts 2-5 investigation cycles per high-risk account), Adaptation (adjusts investigation depth based on red flags and evidence quality), Termination (sufficient evidence obtained, material issue requiring escalation, or maximum cycles reached).

Dimensional profile: HOTL + institutional frame + adaptive goals + stateful.

PCAOB Compliance (Agentic-Specific):

- **AS 1215 (Audit Documentation)**: Workpapers must document *cycle-by-cycle investigation narrative*—what the system perceived, what actions it took, why it adapted—not just final sampling outcome.
- **AS 2315 (Audit Sampling)**: For agentic systems that adaptively investigate, documentation must explain how iteration improved evidence quality and why termination conditions ensured sufficient appropriate audit evidence.
- **Professional Skepticism**: Senior auditor reviews system's investigation logs, assesses escalated accounts, and confirms the system's adaptation logic (adjusting strategy based on discovered evidence) enhances rather than replaces professional judgment.

Governance: Cycle-by-cycle explainability (system generates narrative for each investigation cycle), escalation validation (confirm red-flag accounts appropriately escalated), termination audit (verify sufficient evidence obtained before termination), workpaper integration (investigation logs included in audit documentation). Demonstrates how agentic iteration and adaptation strengthen audit effectiveness when properly governed.

7 Conclusion: Synthesis and Path Forward

Governing agentic systems is not optional—it is the operational prerequisite for deploying these systems responsibly in regulated domains. This chapter has synthesized regulatory obligations, dimensional calibration principles, implementation practices, and organizational accountability structures into a coherent governance framework specific to **agentic systems**—AI systems exhibiting all six GPA+IAT properties (Goal, Perception, Action, Iteration, Adaptation, Termination). This conclusion distills the core imperatives and provides a maturity-based path forward for organizations at different stages of agentic system adoption.

7.1 Three Forces Make Governance Essential

Three converging forces—regulatory momentum, liability exposure, and trust imperatives—make governance a strategic necessity, not compliance theater:

Regulatory Momentum. AI-specific regulation is no longer emerging—it is here. The EU AI Act entered into force in August 2024, establishing enforceable requirements for high-risk AI systems with penalties up to €35 million or 7% of global turnover (European Parliament and Council 2024). U.S. states are enacting their own requirements: Colorado's AI Act (effective January 2026) mandates impact assessments and prohibits algorithmic discrimination (Colorado General Assembly 2024). Sector regulators—Federal Reserve (SR 11-7), PCAOB, SEC, FINRA—are issuing guidance that applies existing standards to AI systems, including agentic systems. The regulatory patchwork is complex and evolving, but the direction is clear: organizations deploying agentic systems in credit, employment, legal, financial, and audit contexts face enforceable obligations. Governance is the mechanism for translating those obligations into operational compliance.

Liability Exposure. Early litigation demonstrates that governance gaps create liability. *Mata v. Avianca* sanctioned an attorney for submitting AI-hallucinated citations—"the AI made the mistake" was not a defense (United States District Court Southern District of New York 2023). ECOA fair lending cases impose strict liability on lenders for disparate impact, regardless of intent or vendor disclaimers. Professional responsibility rules—ABA Model Rules, AICPA standards, fiduciary duties—are non-delegable. Vendor contracts cap liability at subscription fees, shifting risk to deployers. Without governance, organizations face uninsurable, unmitigated risk. With governance—documented risk assessments, monitoring, incident response—organizations create an evidentiary record of reasonable care that may reduce penalties, support litigation defenses, and satisfy regulatory expectations.

Trust and Reputation. Legal, financial, and audit services are trust-intensive. Clients hire attorneys because they trust professional judgment. Investors entrust assets to advisers based on fiduciary obligations. Public companies rely on auditors for independent assurance. Agentic system failures that compromise accuracy, confidentiality, or impartiality erode this trust irreparably. A law firm that discloses client information through an agentic research system's data breach faces not only regulatory sanctions but client defection. An adviser whose agentic financial planning system provides unsuitable recommendations faces not only fiduciary claims but loss of clients. An audit firm whose agentic investigation system produces biased results faces not only PCAOB sanctions but reputational damage. In trust-intensive domains, governance is not merely a legal obligation—it is a competitive necessity.

Governance as Prerequisite, Not Afterthought

Organizations cannot deploy first and govern later. Retrofitting governance onto production systems is costly, disruptive, and often reveals unfixable risks. Governance must be embedded from the outset: risk assessment informs system selection, dimensional calibration guides architecture, logging and monitoring enable accountability, organizational structures assign ownership. This chapter provides the frameworks—regulatory stack, dimensional calibration, implementation controls, accountability models—to build governance into deployment planning.

7.2 Maturity-Based Path Forward

Organizations face agentic system governance from different starting points. We provide maturity-based recommendations:

Organizations Starting from Scratch. If your organization has not yet deployed agentic systems or lacks formal agentic system governance, begin with these foundational steps:

1. **Adopt NIST AI RMF as Baseline**: The NIST AI Risk Management Framework provides flexible, voluntary guidance widely recognized by regulators and industry (National Institute of Standards and Technology 2023). Use its four functions (Govern, Map, Measure, Manage) as your governance scaffold.
2. **Conduct Inventory and Risk Assessment**: Identify all agentic systems currently in use or under consideration (including shadow IT and vendor-provided tools). For each system, verify GPA+IAT properties (Section 2), conduct the dimensional calibration exercise (autonomy, entity frame, goal dynamics, persistence), and perform risk assessment (bias, accuracy, security, privacy, safety, compliance). Prioritize highest-risk agentic systems for immediate governance attention.
3. **Establish Centralized Coordination**: Even if your long-term model is federated or embedded, start with a central AI governance lead or committee to establish policies, build expertise, and prevent inconsistent practices across departments. Centralized governance prevents early-stage chaos.
4. **Focus on Highest-Risk Use Cases First**: Do not attempt to govern all systems simultaneously. Identify the highest-risk deployments—institutional systems with high autonomy, adaptive goals, or access to sensitive data; systems subject to strict regulatory requirements (ECOA, GDPR,

professional ethics)—and implement governance there. Success with high-risk cases builds organizational capability and credibility.

5. **Document Everything**: Even if your governance is basic, document risk assessments, deployment decisions, monitoring results, and incidents. Documentation creates institutional memory, supports audits, and demonstrates good faith to regulators.

Organizations with Partial Governance. If your organization has deployed agentic systems and implemented some governance (e.g., vendor due diligence, basic acceptable use policies), but governance is incomplete or inconsistent, focus on closing gaps:

1. **Audit Against the Five-Layer Framework**: Review your current governance against the five layers from Section 3 (foundational law, professional obligations, sector regulation, AI-specific regulation, voluntary frameworks). Identify gaps: Are you monitoring for ECOA disparate impact? Do your controls satisfy GDPR Article 22 requirements? Have you addressed professional responsibility obligations (ABA, AICPA, fiduciary duty)?
2. **Layer Domain-Specific Controls**: Generic governance frameworks (NIST, ISO) provide structure, but domain-specific requirements (ECOA "principal reasons," PCAOB documentation, attorney confidentiality) require tailored controls. Augment your baseline governance with domain-specific validations, logging requirements, and monitoring procedures.
3. **Formalize Escalation and Accountability (RACI)**: If governance responsibilities are vaguely assigned ("the team is responsible for monitoring"), create a RACI matrix (Table 5). Ensure every governance activity—pre-deployment review, fairness monitoring, incident response—has exactly one accountable party. Test escalation procedures with tabletop exercises.
4. **Implement Continuous Monitoring**: If your governance relies on one-time pre-deployment validation, add continuous monitoring for performance degradation, data drift, concept drift, and fairness violations. Systems validated in 2023 may perform differently on 2024 data; regulatory requirements evolve; adversaries develop new attacks. Governance must be adaptive.
5. **Conduct Post-Incident Reviews**: If incidents have occurred (accuracy failures, user complaints, near-misses), conduct structured post-incident reviews even if no regulatory penalty resulted. Document lessons learned, update risk assessments, and revise controls to prevent recurrence. Incidents are learning opportunities—waste them, and you will repeat them.

Mature Organizations. If your organization has comprehensive agentic system governance—formal policies, dedicated governance teams, continuous monitoring, regular audits—focus on optimization and leadership:

1. **Validate Dimensional Calibration**: Are your controls proportionate to risk? Are you over-governing low-risk systems (creating inefficiency) or under-governing high-risk systems (creating

exposure)? Use Tables 1 through 4 to audit whether control intensity matches system properties.

2. **Participate in Standards Development**: Engage with standards bodies (NIST, ISO, AICPA, ABA), industry groups, and regulatory agencies. Share lessons learned, contribute to best practice development, and influence emerging standards. Mature organizations have governance expertise that benefits the broader community.

3. **Monitor Regulatory Developments Proactively**: Assign personnel to track EU AI Act implementation, U.S. state AI laws, sector regulator guidance, and international developments. Anticipate regulatory changes and adapt governance before enforcement actions occur.

4. **Build Governance as Competitive Advantage**: In trust-intensive domains, demonstrable governance maturity is a market differentiator. Clients, partners, and investors increasingly demand evidence of responsible AI practices. Consider third-party certifications (ISO/IEC 42001), public transparency reports, or governance audits to signal commitment.

7.3 Investing in Governance Capability

Governance is not free. It requires sustained investment across four areas:

Training. Role-based training programs ensure professionals understand AI capabilities, limitations, and governance obligations. Attorneys need training on hallucination risks and Rule 3.3 verification duties. Auditors need training on PCAOB documentation requirements for AI-assisted procedures. Compliance officers need training on fairness metrics and disparate impact analysis. Governance effectiveness depends on organizational competence—budget for initial training and annual refreshers.

Hiring. Governance requires specialized expertise: risk analysts who understand AI fairness testing, compliance officers familiar with ECOA and GDPR, technical specialists who can validate model architectures and explainability techniques, lawyers who understand professional responsibility in AI contexts. Organizations serious about governance must hire or develop these capabilities.

Partnerships. Few organizations can build all governance capabilities in-house. Partner with law firms specializing in AI regulation, consultants offering governance assessments, auditors providing SOC 2 or ISO 42001 certification, and industry groups sharing best practices. Partnerships accelerate capability building and reduce risk of governance blind spots.

Technology. Governance infrastructure—logging systems, monitoring dashboards, fairness testing tools, explainability platforms—requires investment. High-autonomy systems cannot be governed with spreadsheets and manual reviews. Budget for governance tooling as part of AI deployment costs.

7.4 Final Reflection: Governance Enables Sustainable Deployment

Agentic systems offer transformative potential: attorneys can research faster through iterative investigation, advisers can analyze portfolios more comprehensively through adaptive strategy, auditors can investigate anomalies more rigorously through autonomous evidence gathering. But potential is not permission. Deploying agentic systems without governance exposes organizations to regulatory penalties, civil liability, professional discipline, and reputational harm. More fundamentally, it betrays the trust that clients, investors, and the public place in professionals.

Governance is not compliance theater—it is the operational mechanism for maintaining accountability, fulfilling professional duties, and demonstrating that technology serves human objectives rather than displacing human judgment. Done well, governance enables organizations to deploy agentic systems confidently, adapt as risks and regulations evolve, and sustain trust in domains where trust is the foundation of value.

This chapter has provided the conceptual tools: the five-layer regulatory stack, dimensional calibration (mapping GPA+IAT properties to control requirements), implementation controls (iteration auditing, adaptation constraints, termination validation), accountability structures, and worked examples demonstrating how agentic properties create unique governance challenges. The challenge—and opportunity—is to translate these frameworks into your organizational context. The stakes are high, the regulatory landscape is evolving, and the margin for error is narrow. But organizations that invest in agentic system governance today will be positioned to deploy these systems responsibly, defend their practices credibly, and maintain trust durably. That is the path forward.

References

American Bar Association Standing Committee on Ethics and Professional Responsibility (July 2024). *ABA Formal Opinion 512: Generative Artificial Intelligence Tools*. Tech. rep. First comprehensive ABA ethics guidance on generative AI; 15-page opinion addressing competence, confidentiality, communication, candor, supervision, fees. American Bar Association. URL: https://www.americanbar.org/content/dam/aba/administrative/professional_responsibility/ethics-opinions/aba-formal-opinion-512.pdf (visited on 11/21/2024).

Board of Governors of the Federal Reserve System (Apr. 2011). *Supervisory Guidance on Model Risk Management (SR 11-7)*. Tech. rep. Jointly issued with OCC Bulletin 2011-12; adopted by FDIC in FIL-22-2017; establishes framework for model risk management including validation, governance, and ongoing monitoring. Federal Reserve. URL: https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm (visited on 11/21/2024).

Colorado General Assembly (May 2024). *Colorado Artificial Intelligence Act (SB 24-205)*. Senate Bill 24-205. Effective January 2026; mandates impact assessments and prohibits algorithmic discrimination. URL: https://leg.colorado.gov/bills/sb24-205 (visited on 11/21/2024).

Consumer Financial Protection Bureau (2011). *Regulation B (Equal Credit Opportunity)*. 12 CFR Part 1002. Implements ECOA; requires creditors to provide "principal reasons" for adverse credit decisions. URL: https://www.consumerfinance.gov/rules-policy/regulations/1002/ (visited on 11/21/2024).

European Parliament and Council (2016). *General Data Protection Regulation, Article 22: Automated individual decision-making, including profiling*. Regulation (EU) 2016/679, Article 22. Right not to be subject to decisions based solely on automated processing that produce legal or similarly significant effects. URL: https://gdpr-info.eu/art-22-gdpr/ (visited on 11/21/2024).

European Parliament and Council (June 2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Regulation (EU) 2024/1689. Entered into force August 1, 2024; penalties up to €35 million or 7% of global turnover. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689 (visited on 11/21/2024).

International Organization for Standardization (Dec. 2023). *ISO/IEC 42001:2023 Information technology – Artificial intelligence – Management system*. First international AI management system standard; provides requirements for establishing, implementing, maintaining, and continually improving an AI management system. URL: https://www.iso.org/standard/81230.html (visited on 11/21/2024).

National Institute of Standards and Technology (Jan. 2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. Tech. rep. Voluntary framework with four functions: Govern, Map, Measure, Manage; widely recognized by regulators and industry. U.S. Department of Commerce. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (visited on 11/21/2024).

Public Company Accounting Oversight Board (2010a). *AS 1015: Due Professional Care in the Performance of Work*. PCAOB auditing standard on professional skepticism and due care; superseded by AS 1000 effective for audits of fiscal years beginning on or after December 15, 2024. URL: https://pcaobus.org/oversight/standards/auditing-standards (visited on 11/21/2024).

Public Company Accounting Oversight Board (2010b). *AS 1105: Audit Evidence*. PCAOB auditing standard on sufficiency and appropriateness of audit evidence. URL: https://pcaobus.org/oversight/standards/auditing-standards (visited on 11/21/2024).

Public Company Accounting Oversight Board (2010c). *AS 1215: Audit Documentation*. PCAOB auditing standard on workpaper documentation requirements. URL: https://pcaobus.org/oversight/standards/auditing-standards (visited on 11/21/2024).

Public Company Accounting Oversight Board (2010d). *AS 2315: Audit Sampling*. PCAOB auditing standard on statistical and non-statistical sampling. URL: https://pcaobus.org/oversight/standards/auditing-standards (visited on 11/21/2024).

United States Congress (1974). *Equal Credit Opportunity Act*. 15 U.S.C. § 1691 et seq. Prohibits credit discrimination based on race, color, religion, national origin, sex, marital status, age, or receipt of public assistance. URL: https://www.govinfo.gov/content/pkg/USCODE-2021-title15/pdf/USCODE-2021-title15-chap41-subchapIV.pdf (visited on 11/21/2024).

United States District Court Southern District of New York (2023). *Mata v. Avianca, Inc.* No. 22-cv-1461, 2023 WL 4114965 (S.D.N.Y. June 22, 2023). Sanctions of $5,000 each for attorneys who cited six fake ChatGPT-generated cases. URL: https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/ (visited on 11/21/2024).