Agents

Part III: Governing AI Agents

Risk, Compliance, and Accountability in Law and Finance

Jillian Bommarito · Daniel Martin Katz · Michael J Bommarito II

December 12, 2025

Working Draft Chapter

Version 1.01

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 **4**

1 Introduction: The Governance Imperative **4**

1.1 From Tools to Agents: The Governance Shift 4

1.2 The Stakes: Professional Duties Are Non-Delegable 7

1.3 Three Forces Driving Governance Adoption 8

1.4 Mapping Agent Properties to Governance Requirements 9

2 Design Principles: Dimensional Calibration **11**

2.1 The Dimensional Calibration Logic 12

2.2 Autonomy Calibration 13

2.3 Entity Frame Calibration 15

2.4 Goal Dynamics Calibration 17

2.5 Persistence Calibration 20

2.6 Integration: Risk-Calibrated Control Selection 20

3 The Governance Stack: Overlapping Obligations **23**

3.1 The Five-Layer Framework 23

3.2 Layer 1: Foundational Law 25

3.3 Layer 2: Professional and Ethical Obligations 27

3.4 Layer 3: Sector-Specific Regulation 29

3.5 Layer 4: AI-Specific Regulation 30

3.6 Layer 5: Voluntary Governance Frameworks. 32

3.7 Seven Common Controls Across Frameworks 33

4 Implementation: Building Governance Systems **34**

4.1 Risk Assessment as Foundation 35

4.2 Audit Logging: Enabling Reconstruction and Accountability 37

4.3 Explainability: From Technical Outputs to Stakeholder Understanding 40

4.4 Human Oversight: Workflows for HITL, HOTL, and HIC. 42

4.5 Vendor Management: Assessing and Monitoring Third-Party AI 44

4.6 Performance Monitoring and Incident Response 48

5 Accountability and Organizational Structure **52**

5.1 Three Organizational Governance Models 52

5.2 RACI Matrix: Operationalizing Accountability . . . 54
5.3 Escalation and Reporting . . . 55
5.4 Liability Allocation: Who Bears the Risk? . . . 56
6 Examples in Context . . . 59
6.1 Legal Domain: Professional Responsibility and Incident Management . . . 59
6.2 Accounting Domain: Independence and Professional Skepticism . . . 60
7 Conclusion: Synthesis and Path Forward . . . 60
7.1 Three Forces Make Governance Essential . . . 63
7.2 Maturity-Based Path Forward . . . 64
7.3 Investing in Governance Capability . . . 66
7.4 Final Reflection: Governance Enables Sustainable Deployment . . . 67

How to Read This Chapter

This chapter translates the conceptual foundations from Part I and the tactical information from Part II into governance practice. Where Part I asked *what makes a system agentic?* and Part II highlighted *how to construct an agent?*, 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.

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 that *agents* (Level 1) possess these three minimal properties. *Agentic systems* (Levels 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.

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 (*Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023)* 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, whichever is greater.

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 enforcement under the Equal Credit Opportunity Act has traditionally applied disparate impact theory—facially neutral criteria can create liability if they disproportionately harm protected classes without adequate business justification. While the Supreme Court of the United States has not definitively resolved whether ECOA authorizes disparate impact claims, regulators and plaintiffs have long pursued such theories, and prudent lenders treat disparate impact as a material enforcement risk. If an AI credit scoring model produces outcomes that disproportionately harm protected classes, the lender faces significant regulatory and litigation exposure 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*—specifying who may set goals and under what authority, since regulated domains may require approval from compliance officers, general counsel, or clients. Goals must also be *aligned* with actual organizational or client objectives; misaligned goals that optimize for throughput at the expense of quality or minimize cost without considering risk create liability. Furthermore, goals must be *bounded* by constraints that limit aggressive pursuit, preventing agents from ignoring side effects, ethical boundaries, or resource limits. Finally, goals must be *monitorable* so that governance can detect when the agent fails to achieve its objective or when goal pursuit causes unintended harms. This requires the establishment of Key Performance

Indicators (KPIs) and Service Level Agreements (SLAs) that track both 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*—establishing where data comes from, whether it is authoritative, current, and trustworthy, since agents that perceive stale, fabricated, or biased data will make flawed decisions. For third-party systems, establishing provenance can be exceedingly difficult, requiring vendor assessment protocols and documentation of provenance gaps as residual risk. Governance must also address *bias and representation*, determining whether the agent's perceptual model reflects population diversity or encodes historical biases, and implementing bias detection and fairness audits accordingly. *Input validation* is equally critical: adversaries may manipulate what the agent perceives through prompt injection, data poisoning, or adversarial examples, necessitating input validation, sanitization, and anomaly detection. Finally, governance must address *privacy and confidentiality* when perception requires access to sensitive data, ensuring 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 through *action authorization*—defining what actions the agent is permitted to take and requiring explicit authorization for high-consequence actions such as signing contracts or disbursing funds. *Pre-action approval* determines whether certain actions require human approval before execution; human-in-the-loop oversight is appropriate for irreversible or high-stakes actions. Governance must also ensure *rollback and remediation* capabilities: if an action causes harm, can it be undone? Systems must be designed with rollback capabilities and remediation protocols. Finally, *rate limiting* addresses whether the agent can take actions too quickly or too frequently, requiring governance to enforce rate limits and circuit breakers that prevent runaway execution.

Iteration: State Management and Audit Trails. Iteration means the system operates across multiple cycles, each building on prior state. Governance must ensure *reproducibility*—the ability to replay the system's decision sequence, since debugging, auditing, and compliance reviews require reconstructing what the system perceived and why it acted as it did. *State integrity* requires that the system's internal state be protected from tampering or corruption through tamper-evident logging and state validation. Governance must also define *termination conditions* that specify when the system should stop iterating, whether because the goal has been achieved, a resource limit has been reached, or a safety violation has been detected.

Adaptation: Change Control and Revalidation. Adaptation means the system's behavior changes over time. Governance must manage behavioral drift through *change detection*—tracking when the system's behavior changed and what triggered the adaptation, which requires model versioning and

change logs. *Revalidation triggers* determine whether adapted behavior still satisfies safety, fairness, and compliance constraints; governance must define triggers such as performance degradation, distribution shift, or policy updates that initiate revalidation. Governance must also enable *rollback to known-good states* so that if adaptation introduces failures, the system can revert to a prior validated version. Finally, *human oversight of learning* addresses whether adaptation should 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*—the conditions under which the system hands off to a human, such as ambiguous inputs, conflicting objectives, safety violations, or low-confidence decisions. *Graceful shutdown* procedures specify how the system cleanly exits, since abrupt termination may leave systems in inconsistent states. *Handoff procedures* determine what information the system must provide when it escalates to a human, since effective handoff requires context. Finally, *override and emergency stop* mechanisms must allow humans to 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 properties). 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.

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).

These four dimensions were selected because they directly correspond to the technical properties that define agentic behavior (GPA+IAT from Part I) and represent the primary axes along which governance requirements vary across regulatory frameworks.

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: Human-in-the-Loop
Pre-Approval Required

APPROPRIATE WHEN

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

CONTROL REQUIREMENTS

Primary Control: Human reviewer approves every significant action

Logging: Capture recommendation + human decision (approve/reject/modify)

Monitoring: Lighter post-action review (human serves as primary control)

Key Risk: Automation bias—humans rubber-stamping without meaningful review

HITL governance is appropriate when errors carry high consequences and human expertise adds substantial value to decision quality. In legal, financial, and audit contexts, many high-stakes decisions benefit from this mode: the system accelerates research, analysis, or document preparation, but a qualified professional validates the output before it affects clients, counterparties, or regulatory filings. The key governance requirement is ensuring the human review is meaningful—not a rubber stamp—which requires the system to surface sufficient context for informed judgment.

Example: HITL Legal Research

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: Human-on-the-Loop
Monitoring with Intervention

APPROPRIATE WHEN

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

CONTROL REQUIREMENTS

Primary Control: Monitoring dashboards with anomaly detection

Logging: Detailed—enables retrospective audit of decisions not reviewed prospectively

Escalation: Triggers for low-confidence decisions, policy boundary cases, user complaints

Intervention: Protocols for human override when anomalies detected

HOTL governance suits situations where transaction volume makes individual review impractical, but errors remain detectable and correctable through monitoring. The system handles routine operations autonomously while humans watch for anomalies—unusual patterns, error spikes, or edge cases that exceed the system's training distribution. Governance focuses on defining clear escalation triggers, maintaining real-time dashboards, and ensuring intervention mechanisms actually work when needed.

Example: HOTL Customer Service

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: Human-in-Command
Strategic Oversight and Emergency Stop

APPROPRIATE WHEN

- ▶ System operates at scale and speed that precludes individual review (millions of transactions daily)
- ▶ Actions are individually low-consequence but cumulatively significant
- ▶ System operates in a stable, well-understood environment with strong safeguards

CONTROL REQUIREMENTS

Primary Control: Comprehensive logging + statistical monitoring

Logging: Exceptionally detailed—post-action auditability is critical

Monitoring: Statistical (fairness metrics, error trends, drift detection)

Emergency Stop: Accessible to authorized personnel; tested regularly

The Autonomy-Auditability Trade-off. As summarized in Table 1, 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."

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. Table 2 summarizes entity frame calibration.

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.

Example: HIC Fraud Detection

A credit card 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.

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

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: Human Entity Frame

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: Hybrid Entity Frame

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: Machine Entity Frame

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: Institutional Entity Frame

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.

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. Table 3 summarizes goal dynamics calibration.

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

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

whether the system achieves the goal and whether side effects emerge.

Example: Static Goals

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.

Adaptive goals can drift within their permitted boundaries, and subtle drift may escape notice until it causes harm. Detecting drift requires both a clear standard and active surveillance. The standard comes from explicitly defining which aspects of the goal may adapt and which constraints are inviolable—without this distinction, there is nothing to measure against. Surveillance comes from a monitoring framework that specifies review frequency and revalidation triggers, catching deviation before it compounds. Detection alone is insufficient, however; organizations also need rollback capability, enabling the system to revert to a validated state when adaptation degrades performance or violates constraints.

Example: Adaptive Goals

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. When a system proposes changes to its own objectives, every proposal requires human validation before implementation—the system cannot autonomously alter its goals. This creates the highest governance burden of any goal dynamics category. Organizations must designate approval authority, typically reserving this role for senior leadership or a governance committee given the stakes involved. The approval process demands rigorous documentation: why is the system proposing this change, what evidence supports it, and what are the potential consequences? Even after approval, governance is not complete; the modified system must undergo revalidation to confirm it remains safe, fair, and compliant. This cumulative overhead explains why negotiated goals appear only in the most sophisticated agentic deployments.

Example: Negotiated Goals

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: Goal Dynamics Calibration: Objective Stability and Control Requirements

Goal Dynamics	Description	Example Use Cases	Control 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

2.5 Persistence Calibration

Persistence determines whether the system maintains state across interactions. Stateful systems create compounding error risk and require state integrity controls. Table 4 summarizes persistence calibration.

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 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: Stateless System

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. Stateful systems accumulate information across interactions, meaning each decision builds on prior context. This creates a distinctive governance challenge: errors compound. A misunderstanding in one session can propagate through subsequent decisions, corrupting an entire chain of reasoning. Governance must therefore protect state integrity against tampering, corruption, and adversarial manipulation. Comprehensive logging is essential to enable reconstruction of how state evolved—without it, diagnosing problems becomes nearly impossible. Monitoring must specifically watch for error compounding, catching cases where an early mistake ripples through later decisions. Organizations also need clear criteria for state resets: when should the system discard accumulated context and start fresh? Common triggers include user logout, policy changes, and detected anomalies, though the right criteria depend on deployment context.

Example: Stateful System

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.

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.

Table 4: Persistence Calibration: State Management and Control Requirements

Persistence	Description	Example Use Cases	Control 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

Controls must respond to this multidimensional risk.

Table 5 compares two contrasting risk profiles. The **low-risk** example is a legal research assistant that verifies whether citations in a brief exist in official reporters and accurately support the propositions for which they are cited. An attorney reviews every output before it reaches a client or court, and each verification stands alone without reference to prior work. The **high-risk** example is a bank's mortgage underwriting system that evaluates applications, requests documentation, and renders credit decisions. The system learns from outcomes over time, tracks applicant history across multiple interactions, and operates at scale with humans reviewing aggregate performance metrics rather than individual decisions. The low-risk profile demonstrates how strong human oversight compensates for other dimensions, enabling lighter governance. The high-risk profile shows how compounding risk across all four dimensions necessitates intensive controls.

Table 5: Risk Profile Comparison: Low-Risk vs. High-Risk Agentic Systems

		Low-Risk: Legal Research	High-Risk: Credit Underwriting
Dimensional	Autonomy	HITL—attorney reviews all outputs	HIC—autonomous decisions; aggregate monitoring
	Entity Frame	Human—operates under attorney's name	Institutional—acts on behalf of bank
	Goal Dynamics	Static—fixed citation verification goal	Adaptive—adjusts feature weights from feedback
	Persistence	Stateless—each verification independent	Stateful—maintains applicant history
Control Calibration	Logging	Basic inputs/outputs (attorney review is primary control)	Comprehensive—inputs, model version, weights, rationale, state changes
	Monitoring	Minimal (quarterly spot-checks)	Continuous—monthly fairness metrics, drift detection
	Explainability	Not required (attorney verifies)	ECOA-compliant "principal reasons"
	Fairness	Not applicable	Pre-deployment + continuous + revalidation
	Vendor Mgmt	Standard SaaS due diligence	Enhanced—interpretability, audit rights
	Human Oversight	Attorney review per output	Monthly, quarterly, annual board review
	Incident Response	Standard error correction	Immediate halt; regulator notification

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

Having established dimensional calibration principles in Section 2, we now map the regulatory landscape those controls must satisfy. 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 legal obligations governing data protection, discrimination, consumer protection, and contracts. This layer encompasses not only statutes—such as the General Data Protection Regulation (GDPR), Equal Credit Opportunity Act (ECOA), and state consumer protection statutes—but also tort law as well as other common law principles. Common law doctrines provide protections that predate and operate independently of statutory schemes and include:

   - Defamation law constrains AI-generated content that makes false statements of fact about identifiable individuals.
   - Negligence principles may impose duties of care when deploying systems that foreseeably cause harm.
   - Privacy torts (intrusion upon seclusion, public disclosure of private facts) create liability independent of data protection statutes.

   These legal foundations 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.

Table 6: Summary of the Five-Layer Governance Framework

Layer	Source	Key Examples
1. Foundational Law	Statutes, regulations, common law	GDPR, ECOA, Tort law
2. Professional Obligations	Bar associations, regulatory bodies	ABA Model Rules, fiduciary duty, AICPA standards
3. Sector-Specific Regulation	Industry regulators	SR 11-7, FINRA rules, PCAOB standards
4. AI-Specific Regulation	AI-targeted laws	EU AI Act, Colorado AI Act, NYC Local Law 144
5. Voluntary Frameworks	Standards bodies, industry groups	NIST AI RMF, ISO/IEC 42001, SOC 2

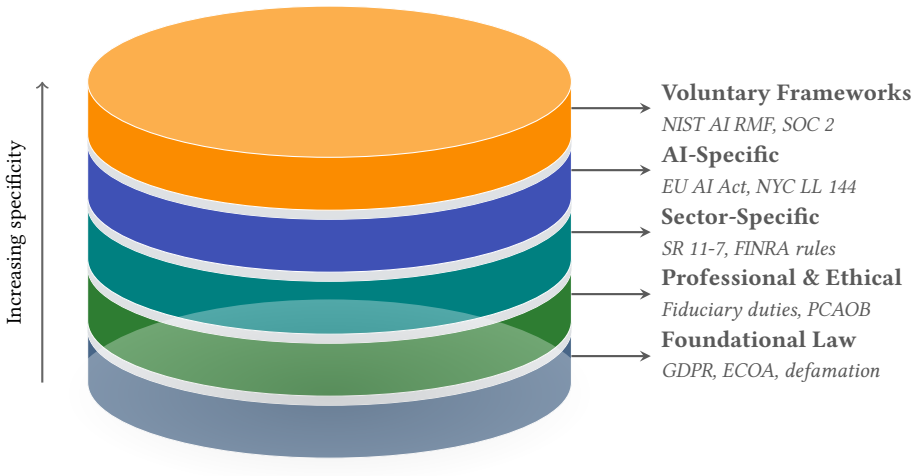


Figure 1: The five-layer governance framework as stacked obligations. Foundational law sets the baseline; professional duties add licensed-practice requirements; sector-specific rules address industry contexts; AI-specific regulation targets algorithmic systems; and voluntary frameworks supply certification and best-practice guidance. All applicable layers must be addressed—no single layer provides complete coverage.

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, and 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 commonly used in disparate impact analysis.

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 (*Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023)* 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 en-

forcement, 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 (SB 24-205)**: 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 (monitor-

ing 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.

The system 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. *Dimensional profile: HITL + hybrid frame + adaptive goals + stateful.*

The risk assessment identifies five primary concerns, summarized in Table 7.

Compliance risk ranks highest: the system may recommend unsuitable investments, violating Advisers

Act fiduciary duty. Unlike a simple Q&A chatbot, iteration and adaptation create compounding risk—a flawed recommendation in cycle one shapes subsequent analysis.

Accuracy risk stems from potential hallucination of market data or misinterpretation of client constraints. Across iterative cycles, these errors compound rather than self-correct.

Adaptation risk arises when the system drifts from regulatory compliance—for example, learning to recommend higher-fee products based on firm incentives rather than client best interest.

Iteration risk manifests as either excessive iteration (analysis paralysis) or premature termination (incomplete analysis).

Security risk, though less likely, carries the highest impact: prompt injection across iterative cycles could manipulate accumulated state, potentially disclosing other clients' information.

Table 7: Risk Assessment: Agentic Financial Planning Assistant

Risk Category	Likelihood	Impact	Key Mitigations	Residual Risk
Compliance	High	High	HITL approval before client recommendations; monthly compliance review; quarterly fiduciary assessment	Moderate
Accuracy	Moderate	High	Verified data sources (Bloomberg, Reuters); cross-cycle consistency checks; human review of final recommendations	Moderate
Adaptation	Moderate	High	Adaptation limited to analysis methods; recommendation criteria fixed; quarterly adaptation audit	Low - Moderate
Iteration	Low	High	Explicit termination conditions (5 cycles OR conf. >0.85 OR 14-day timeout); human review on timeout	Low
Security	Low	Very High	Input sanitization per cycle; state integrity validation; client data isolation; monthly penetration testing	Low

Monitoring: Daily compliance review, weekly adaptation log review, monthly accuracy and termination analysis, continuous client feedback.

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), and analyzing trends across multiple applicants. The system adapts its investigation strategy based on discovered risk patterns and terminates when sufficient information is gathered or escalation is required.

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

The system's agentic properties shape its governance requirements. Its **goal** is to approve qualified applicants while managing credit risk and satisfying ECOA requirements. Its **perception** draws on application data, third-party verification responses, and historical default patterns. Its **actions** include requesting documents, querying APIs, and generating 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** occurs under explicit conditions: confidence exceeds 0.90, the maximum of 7 cycles is reached, or red flags trigger escalation to a 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 are encrypted at rest, with access restricted to compliance officers, auditors, and authorized investigators. The system uses append-only storage with cryptographic integrity verification (tamper-evident) and 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—analyzing transactions, requesting documentation, cross-referencing third-party data, adapting its strategy based on discovered anomalies, and escalating to senior auditors when material issues arise. *Dimensional profile: HOTL + institutional frame + adaptive goals + stateful.*

PCAOB Auditing Standards require auditors to document the rationale for procedures in workpapers (Public Company Accounting Oversight Board 2010c; Public Company Accounting Oversight Board 2010d). For agentic systems, explainability must capture *why* the system escalated certain accounts, *how* its strategy evolved across cycles, and *what* evidence supported termination decisions. Figure 2 illustrates the workflow. Each cycle generates explanations logged to audit 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

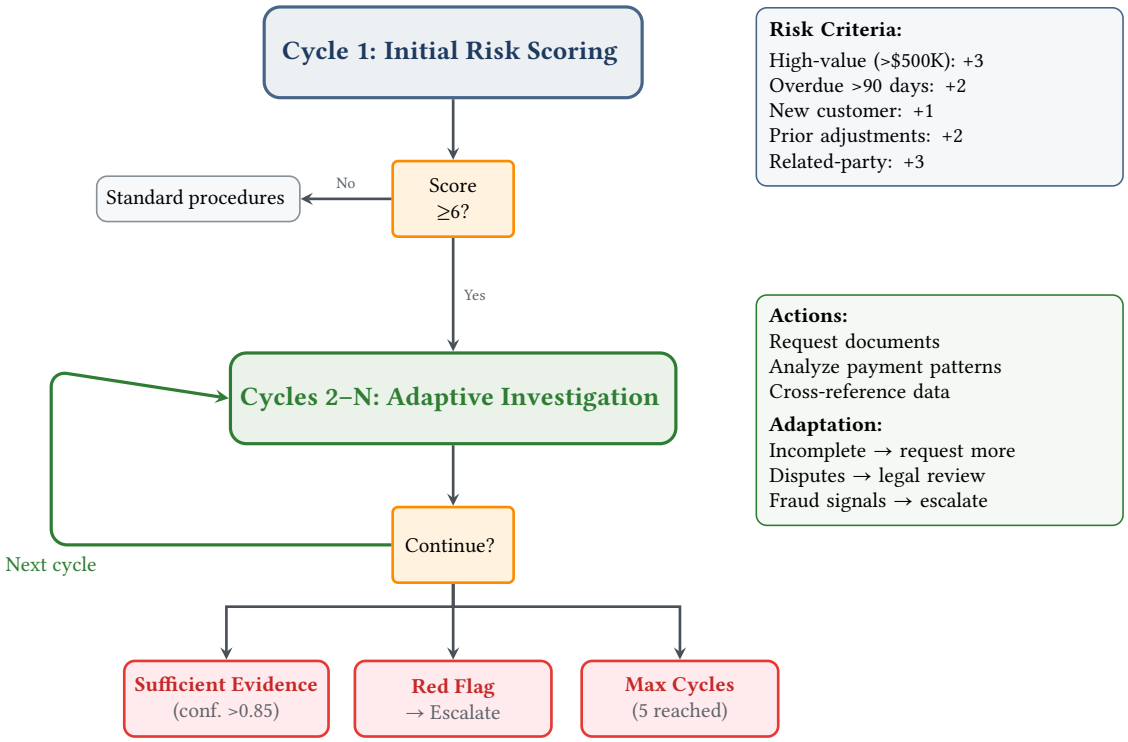


Figure 2: Iterative investigation workflow with explainable adaptation. Cycle 1 scores accounts using risk criteria; high-risk accounts (score ≥6) proceed to iterative investigation. Cycles 2–N gather evidence and adapt strategy based on findings. Investigation terminates when sufficient evidence is obtained, a red flag requires escalation, or maximum cycles are reached.

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 through a dashboard (Figure 3) displaying agentic-specific metrics and escalation triggers. When escalation frequency spikes or average cycles increase significantly, supervisors investigate root causes—data quality degradation, overly conservative termination thresholds, or emerging risk patterns requiring strategy adjustment.

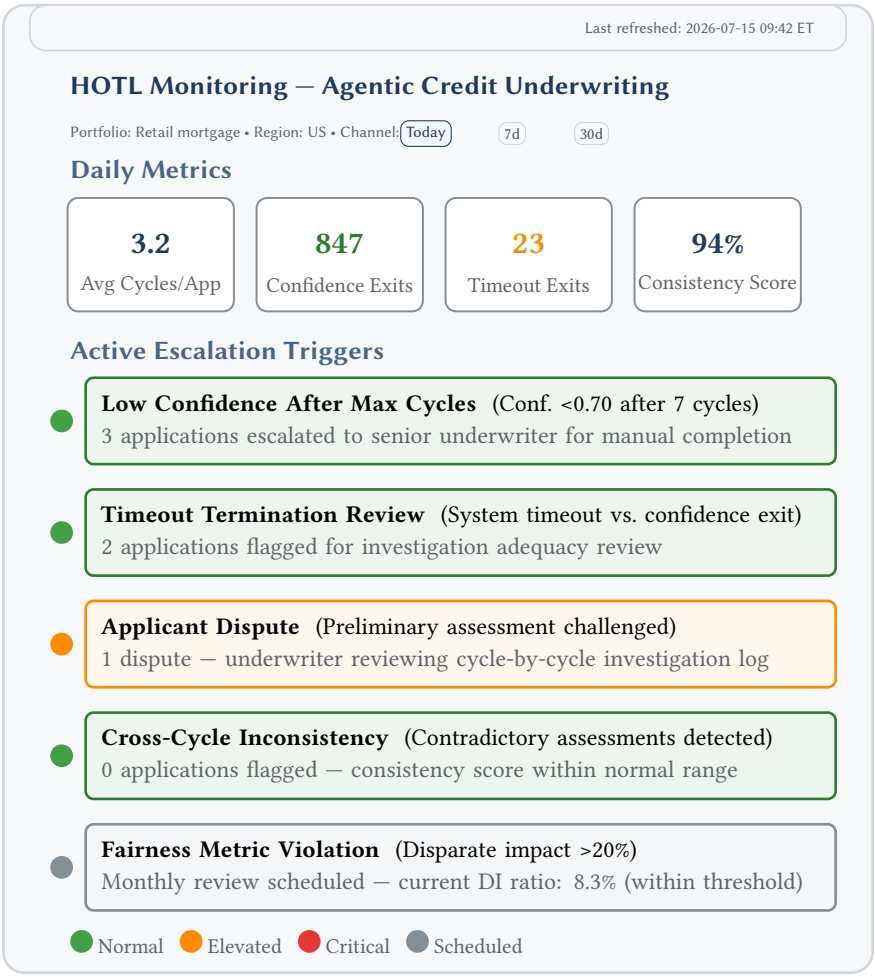


Figure 3: HOTL monitoring dashboard for agentic credit underwriting. The dashboard displays daily operational metrics (average investigation cycles, termination reasons, cross-cycle consistency) and active escalation triggers with status indicators. Senior underwriters monitor aggregate performance and intervene when triggers fire.

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). The framework proceeds through three phases. **Phase 1 (Assessment)** uses questionnaires to gather information about the vendor's data sources, model architecture, security practices, performance benchmarks, and fairness testing methodology. **Phase 2 (Document Review)** examines supporting documentation—SOC 2 reports, data processing agreements, model validation reports, and security certifications—to verify vendor claims. **Phase 3 (Validation)** confirms claims through reference checks with existing clients in similar domains and pilot testing with representative data to validate accuracy, explainability, and performance before full deployment.

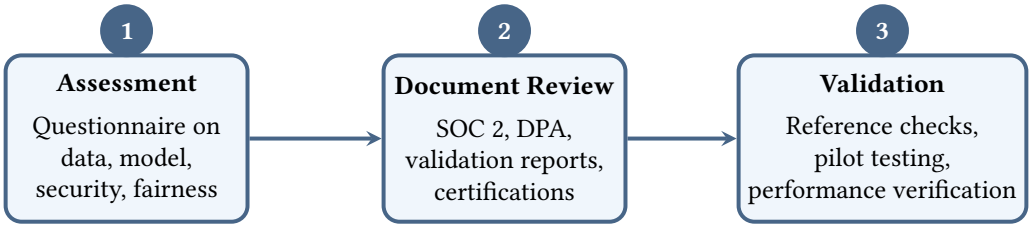


Figure 4: Three-phase vendor due diligence framework. Phase 1 gathers information through questionnaires; Phase 2 reviews supporting documentation and certifications; Phase 3 validates claims through reference checks and pilot testing with representative data.

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

Contract Term	Negotiation Focus
Liability Caps	Negotiate higher caps or uncapped liability for confidentiality breaches and gross negligence in high-risk use cases (credit decisioning, legal advice, audit).
Model Update Notification	Require 30-60 days advance notice before material model updates, enabling revalidation 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 termination; specify jurisdiction for storage.
SLAs	Define performance thresholds (accuracy, uptime, response time) and specify remedies for violations.

Table 8: Key contract terms for allocating risk in AI vendor agreements.

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 contractual provisions that address adaptation opacity:

Clause Type	Purpose
Adaptation Mechanism Disclosure	Vendor identifies all adaptation mechanisms before contract execution, specifying which components adapt and which remain static.
Change Log Access	Customer receives API or dashboard access to logs showing what changed, when, and why.
Material Change Thresholds	Define triggers (feature weight shifts, performance degradation, fairness drift) requiring vendor notification and customer revalidation rights.
Audit and Testing Rights	Customer may conduct periodic behavioral validation testing; vendor cooperates with third-party audits.
Adaptation Freeze Options	Customer may temporarily disable learning mechanisms during regulatory examinations or incident investigations.

Table 9: Five categories of adaptation transparency clauses for agentic vendor contracts.

Example Contractual Language: Adaptation Transparency

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.

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. This shift from inspection-based to continuous monitoring mirrors the evolution of Statistical Process Control in manufacturing—a historical parallel we examined in Part II.

Figure 5: The seven-stage incident response cycle for AI system failures. The dashed feedback loop emphasizes that post-incident review improves detection capabilities, creating continuous improvement in governance controls.

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. Regulatory expectations and enforcement practice under ECOA effectively require lenders to monitor for disparate impact as part of fair lending compliance. Similarly, GDPR Article 22 requires ongoing assessment of automated decision-making.

Example: Disparate Impact in Agentic Credit Underwriting. A regional bank deploys an agentic mortgage underwriting system that iteratively investigates applicants: it performs initial risk scoring, requests documentation based on risk indicators, adapts its investigation strategy based on applicant responses, and terminates when sufficient information is gathered or confidence thresholds trigger human escalation. *Dimensional profile: HIC + institutional frame + adaptive goals + stateful.*

During routine monthly fairness monitoring, the compliance team detects a significant disparity: Hispanic applicants have a 65% approval rate compared to 82% for white applicants—a clear violation of the 80% rule commonly used in disparate impact analysis (65/82 = 79.3%). Following the Tier 3 escalation pathway for critical issues (see Figure 9), the compliance analyst immediately notifies the Chief Risk Officer, who halts the system and escalates to the CEO within hours of detection (Figure 6).

Ten weeks later, with the investigation complete and remediation implemented, the CRO sends a closure report (Figure 7). The investigation reveals a governance challenge unique to agentic systems: the discrimination did not originate in the scoring model itself, which was facially neutral and passed traditional fairness testing. Instead, bias emerged through *how the system investigated applicants across cycles*—Hispanic applicants triggered more verification cycles, leading to higher abandonment rates before final decisions were rendered. The system was discriminating in its *process*, not its *decisions*.

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) is insufficient; organizations must audit investigation process across cycles.
- **Adaptation constraints**: System 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.

From: Sarah Chen, Chief Risk Officer <s.chen@firstnational.com> URGENT
To: Michael Torres, Chief Executive Officer <m.torres@firstnational.com>
Cc: Jennifer Walsh, General Counsel; Board Risk Committee Chair
Date: June 15, 2024, 3:12 PM
Subject: **URGENT: Fair Lending Violation Detected – Agentic Underwriting System Halted**

Michael,

I'm escalating a critical issue that requires your immediate awareness. Our compliance team's monthly fairness review identified a significant disparity in the agentic mortgage underwriting system this morning: Hispanic applicants have a 65% approval rate compared to 82% for white applicants. This violates the 80% rule commonly used in disparate impact analysis (65/82 = 79.3%) and constitutes a potential ECOA fair lending violation.

Actions taken in the last two hours:

- Halted the agentic underwriting system immediately per emergency stop protocol
- Reverted all mortgage applications to manual underwriting
- Preserved 12 months of cycle-level system logs in read-only archive
- Convened incident response team: Jennifer Walsh (GC) is leading legal strategy, Marcus Thompson (AI Governance), and Rita Gonzalez (Mortgage Systems Lead)

I just got off the phone with Jennifer. As you know, ECOA enforcement doesn't require proof of discriminatory intent—regulators and courts have applied disparate impact analysis, meaning the effect matters regardless of motive. She's already reviewing our regulatory notification obligations and will have a timeline for CFPB self-reporting by tomorrow. We face potential enforcement action, class action exposure, and reputational risk.

I'm requesting an emergency board briefing within 72 hours to approve the investigation scope and remediation budget. I'll send a detailed findings report once the investigation is complete.

Sarah

Sarah Chen | Chief Risk Officer
First National Bank | Tel: (555) 234-5678
s.chen@firstnational.com

This email and any attachments are confidential and intended solely for the addressee. If you have received this message in error, please notify the sender immediately and delete this message. Unauthorized use, disclosure, or copying is prohibited.

Figure 6: Initial escalation email from the Chief Risk Officer to CEO upon detection of a fair lending violation. The email documents immediate containment actions taken within two hours of detection, consistent with the Tier 3 escalation pathway for critical issues.

From: Sarah Chen, Chief Risk Officer <s.chen@firstnational.com>
To: Michael Torres, Chief Executive Officer <m.torres@firstnational.com>
Cc: Jennifer Walsh, General Counsel; Board Risk Committee Chair
Date: August 30, 2024, 4:47 PM
Subject: **CLOSED: Fair Lending Incident FL-2024-003 – Findings and Remediation**

Michael,

I'm pleased to confirm that Incident FL-2024-003 is now closed. The system has been remediated and redeployed. Here's what we found and what we've changed.

Root Cause: Process-Based Discrimination. The discrimination wasn't in the scoring model—Cycle 1 risk assessment was facially neutral and passed our traditional fairness tests. The problem was *how the system investigated applicants across subsequent cycles*:

- Hispanic applicants triggered more verification cycles (5.2 avg vs. 3.8 for white applicants)
- The system had learned to flag shorter U.S. employment tenure for extra scrutiny—a proxy for national origin
- Extended verification led to higher abandonment: 28% of Hispanic applicants withdrew vs. 12% of white applicants

The system wasn't discriminating in its *decisions*—it was discriminating in its *process*. Applicants who would have been approved were dropping out because we were investigating them more aggressively.

Remediation Implemented. Rita Gonzalez and Marcus Thompson led the technical remediation. Before redeploying, we made four changes: (1) prohibited employment tenure from influencing investigation depth; (2) added cycle-count parity monitoring that flags deviations >20% from demographic medians; (3) implemented abandonment tracking by protected class; and (4) retrained and revalidated the system for both outcome *and* process fairness.

Regulatory and Applicant Notification. Jennifer filed our self-report with the CFPB on July 12th, within the 30-day window the board approved. Her team has also sent notification letters to all 847 affected Hispanic applicants from the past 12 months, offering expedited re-review with human oversight, waived fees, and priority processing. So far, 312 have requested re-review.

What We're Changing Going Forward. This one caught us off guard. Our fairness testing was focused on approval rates—we weren't looking at how the system *investigated* people differently. I talked with Marcus yesterday, and he's already updating our AI Governance framework to treat "iteration bias" as a distinct risk category. We're revising our validation protocols to require fairness testing across the full investigation process, not just final decisions. We're also adding this to the quarterly model risk review.

I'll present the full post-incident review at next month's board meeting, including proposed policy updates for all agentic systems.

Sarah

Sarah Chen | Chief Risk Officer
First National Bank | Tel: (555) 234-5678
s.chen@firstnational.com

Figure 7: Closure email summarizing investigation findings and remediation actions. The root cause—process-based discrimination through unequal verification burdens—represents a failure mode unique to agentic systems that traditional outcome-focused fairness testing would not detect.

These technical and operational controls require clear organizational ownership and accountability structures—the subject we turn to next.

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). This office establishes policies, reviews all proposed AI deployments, conducts risk assessments, and monitors 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 requirements (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. Each function applies its existing 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).

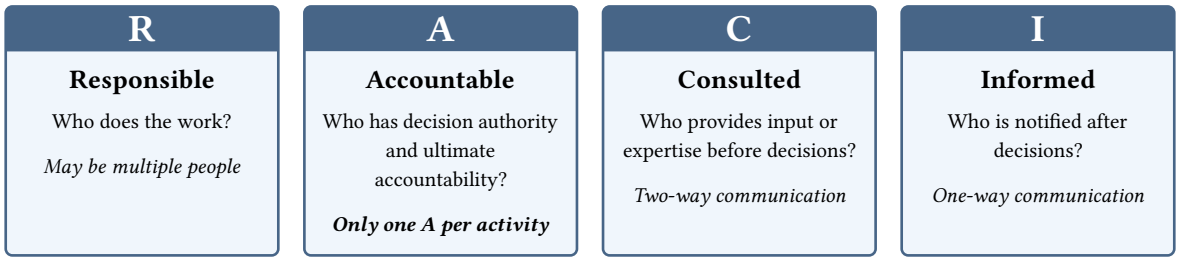
Accountability ambiguity: If AI governance is "everyone's responsibility," it may become no one's priority.

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:



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 10 provides a sample RACI matrix for AI governance activities.

Table 10: Sample RACI Matrix for AI Governance Activities

Activity	Board / CEO	CRO / CCO	AI Gov. 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	R

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 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.

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.

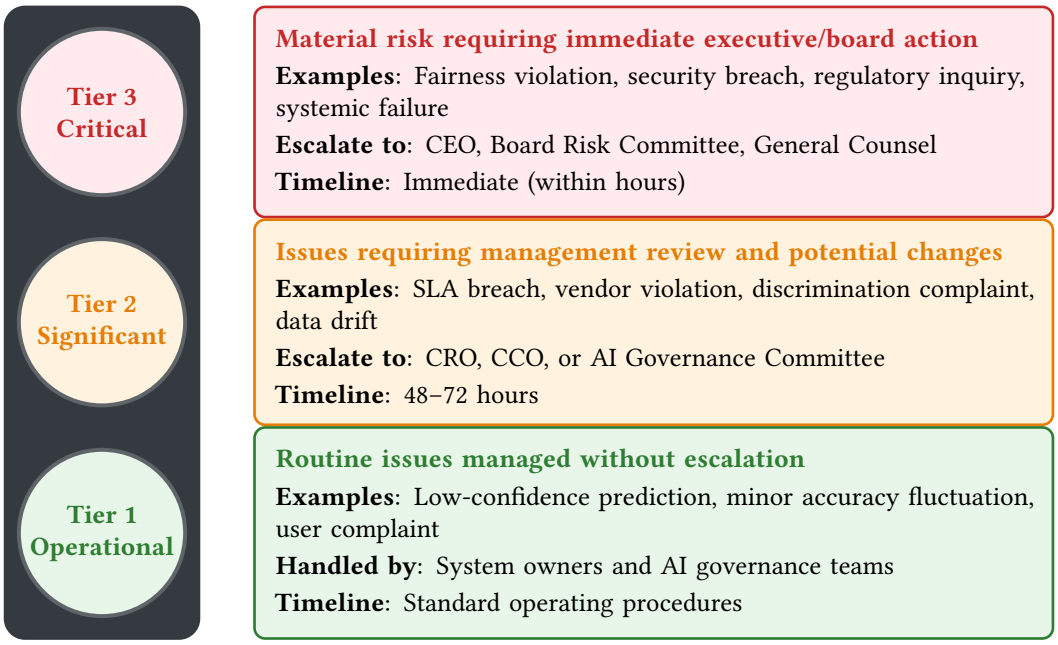


Figure 8: Three-tier escalation model for AI governance issues. Tier 1 (operational) issues are handled routinely; Tier 2 (significant) issues require management review; Tier 3 (critical) issues demand immediate executive or board action. Pre-defining which issues fall into each tier ensures rapid, consistent response.

Example Escalation: Fairness Violation in Credit Decisioning. A bank's monthly fairness monitoring detects disparate impact in credit pre-screening (see Section 4.6). Figure 9 illustrates the Tier 1 → Tier 3 escalation pathway, demonstrating how pre-defined critical issues trigger rapid organizational response with specific time targets at each stage.

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.

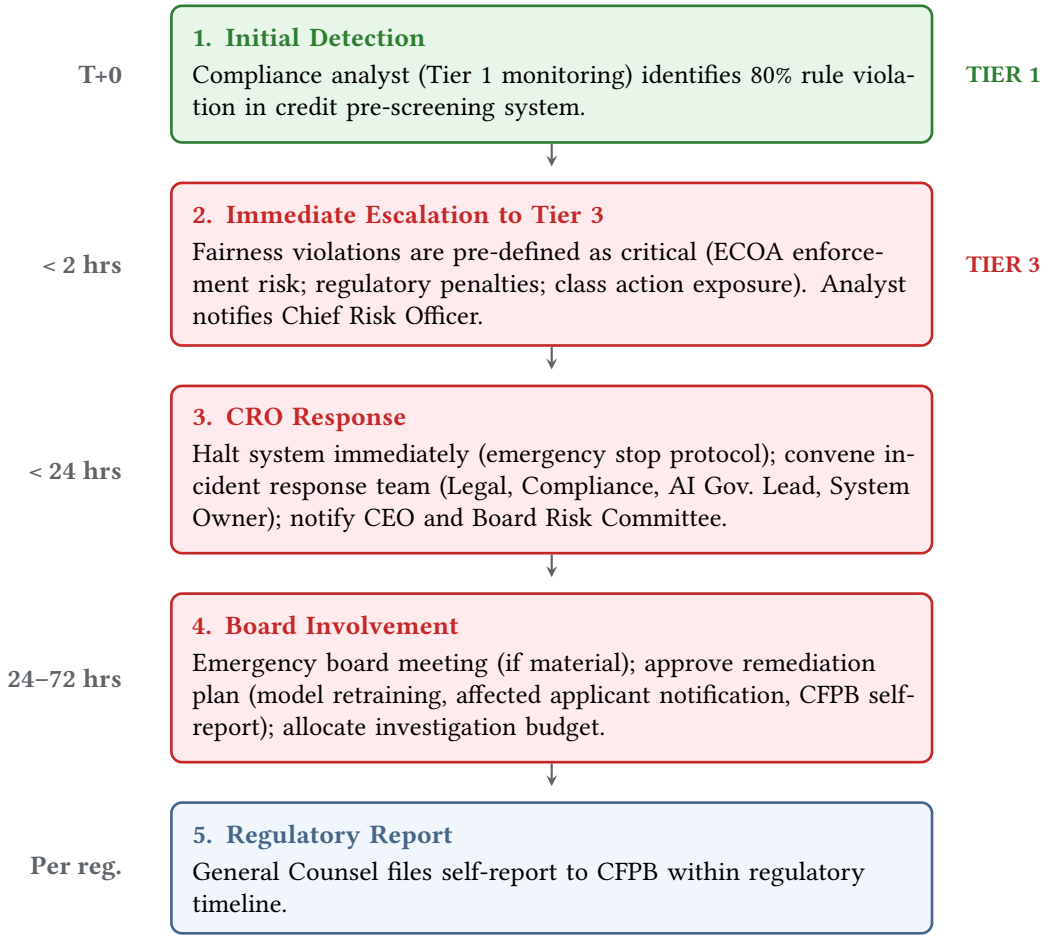


Figure 9: Example escalation pathway for a fairness violation detected in credit decisioning. The pathway demonstrates rapid Tier 1 to Tier 3 escalation for pre-defined critical issues, with specific time targets at each stage ensuring prompt organizational response.

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**: Under traditional enforcement practice, ECOA liability attaches to 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 and accounting contexts. (Financial services examples—including credit underwriting, financial planning, and fair lending compliance—are developed throughout Section 4.) 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. *Dimensional profile: HITL + human frame + static goals + stateless.*

Figure 10 documents an incident where the system's cross-cycle adaptation introduced citation errors that propagated through subsequent iterations—a failure mode unique to agentic systems. The incident report follows ISO 27001 incident management standards while illustrating three governance lessons. First, iteration and adaptation compound errors across cycles, making single-point output review insufficient. Second, confidence thresholds must incorporate domain-specific accuracy metrics, not just relevance scores. Third, professional duty under Rule 1.1 requires attorneys to understand iterative system logic, not merely review final outputs.

6.2 Accounting Domain: Independence and Professional Skepticism

Example 2: 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.*

Figure 11 shows an excerpt from the firm's policy. The policy establishes guiding principles (independence, competence, confidentiality), distinguishes permitted uses (research, analytics, documentation assistance) from prohibited uses (management decisions, audit opinions, unauthorized data sharing), and implements safeguards through vendor approval requirements, mandatory professional review, and documentation standards. Training requirements ensure personnel understand both tool capabilities and professional obligations. Incident reporting procedures establish clear escalation pathways when independence concerns or data breaches arise.

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 (management decisions, audit opinions) uses.
- **Accountability assignment**: Partners responsible for reviewing AI-assisted work; National Office Ethics Group accountable for policy updates.

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

MORRISON & STERLING LLP
Risk Management — Incident Response

Incident No. IR-2024-017
Status: CLOSED

AI System Incident Report

Agentic Legal Research Assistant — Cross-Cycle Hallucination

Detected: Mar 15, 2024 **Severity:** High **Category:** Output Accuracy
Closed: Apr 12, 2024 **Reporter:** Opposing Counsel **Owner:** Chief Risk Officer

1. INCIDENT DESCRIPTION

Opposing counsel in *Martinez v. Consolidated Industries* alerted the firm that a summary judgment motion contained citations that, while identifying real cases, mischaracterized holdings. The motion was prepared using the firm's agentic legal research system, which iteratively investigates legal questions through 2–6 research cycles, adapting search strategies based on relevance patterns.

2. IMPACT ASSESSMENT

2.1 *Professional Responsibility.* Potential violation of ABA Model Rule 3.3 (Candor Toward the Tribunal) due to submission of inaccurate case characterizations.

2.2 *Client Impact.* Motion credibility compromised; client notified; fee reduction offered.

2.3 *Scope.* Review of 47 prior research sessions identified 6 additional sessions (13%) with cross-cycle adaptation errors.

3. IMMEDIATE RESPONSE

3.1 System access suspended firm-wide pending investigation (Mar 15).

3.2 Corrected motion filed with court; candor-to-tribunal explanation submitted per Rule 3.3 (Mar 16).

3.3 Client notified of incident and offered fee reduction for affected matter (Mar 17).

3.4 All pending matters using system flagged for manual citation verification (Mar 17).

4. ROOT CAUSE ANALYSIS

4.1 *Cycle-Level Audit.* Investigation of iteration logs revealed: Cycles 1–3 correctly identified 8 relevant cases. Cycle 4 detected contradictory authority and attempted to "harmonize" holdings through paraphrasing—introducing mischaracterization. Cycles 5–6 propagated the erroneous synthesis without detecting the error.

4.2 *Adaptation Failure.* The system's contradiction-resolution logic created hallucination risk by paraphrasing holdings rather than preserving verbatim quotations.

4.3 *Termination Failure.* System terminated based on confidence threshold (>0.85) despite holding mischaracterization; confidence metric measured legal relevance but did not capture citation accuracy.

5. CORRECTIVE ACTIONS

5.1 *Adaptation Constraints.* System reconfigured to prohibit paraphrasing of holdings; require verbatim quotations with Bluebook pin cites for all case references.

5.2 *Cross-Cycle Consistency.* Implemented automated flagging when later cycles contradict earlier findings; contradictions now escalate to attorney rather than automated resolution.

5.3 *Termination Revision.* Confidence threshold revised: system terminates only when legal relevance confidence >0.85 AND citation accuracy score >0.95 (verified via database cross-check).

5.4 *HITL Verification.* Attorneys must review cycle-by-cycle logs, not just final output; research workpapers must document validation of each cited case.

6. PREVENTIVE ACTIONS

6.1 Quarterly iteration audits established: sample 15% of research sessions; review cross-cycle adaptation patterns for citation accuracy.

6.2 Training updated: all attorneys complete 2-hour module on agentic system risks and cycle-level review requirements.

6.3 Vendor notified of defect; contractual SLA for accuracy monitoring invoked.

7. LESSONS LEARNED

7.1 Iteration and adaptation compound errors across cycles; single-point output review is insufficient for agentic systems.

7.2 Confidence thresholds must incorporate domain-specific accuracy metrics (citation fidelity), not just relevance scores.

7.3 Professional duty (Rule 1.1 competence) requires attorneys to understand iterative system logic, not merely review outputs.

RELATED INCIDENTS: None **REVIEW CYCLE:** 90 days (Jul 12, 2024)

Figure 10: Incident report for an agentic legal research system failure. The report follows ISO 27001 incident management standards (classification, root cause analysis, corrective and preventive actions) while documenting an agentic-specific failure mode: cross-cycle error propagation where the system's adaptation logic introduced hallucinations that compounded across subsequent iterations.

BIG FOUR ACCOUNTING LLP
National Office — Professional Standards

Policy No. AI-2024-003
Version 2.1

Artificial Intelligence Acceptable Use Policy

Audit, Tax, and Advisory Services

Owner: Chief Ethics Officer **Approved by:** Executive Committee **Effective:** Jan 1, 2025 **Review by:** Jan 1, 2026

1. SCOPE

This policy governs the use of artificial intelligence tools, including large language models, agentic systems, and automated analytics platforms, in all audit, tax, and advisory engagements conducted by firm personnel.

2. DEFINITIONS

2.1 *AI Tool.* Any software system that uses machine learning, natural language processing, or automated reasoning to analyze data or generate outputs.

2.2 *Agentic System.* An AI tool that operates iteratively, adapts its behavior based on feedback, and determines when to escalate to human review.

2.3 *Professional Review.* Evaluation by a manager or partner to assess accuracy, completeness, and appropriateness of AI-generated work product.

3. GUIDING PRINCIPLES

3.1 *Independence.* AI tools shall not be used in any manner that impairs auditor independence under AICPA Professional Standards or SEC Rule 2-01. Personnel shall not delegate management decisions to AI systems or use AI outputs that create self-review threats.

3.2 *Professional Competence.* Personnel using AI tools must understand the tool's capabilities, limitations, and potential for error. Use of AI does not diminish the professional's responsibility to exercise due care under AT-C Section 105.

3.3 *Confidentiality.* Client data processed by AI tools must be protected consistent with firm confidentiality obligations and applicable data processing agreements.

4. PERMITTED USES

AI tools may be used for: (a) research and analysis, including accounting standards research, industry benchmarking, and regulatory guidance review; (b) data analytics, including anomaly detection, transaction testing, and sampling optimization; and (c) documentation assistance, including workpaper drafting and memo summarization, subject to professional review requirements in Section 6.

5. PROHIBITED USES

The following uses are prohibited without exception:

5.1 Using AI to make or recommend management decisions for audit clients.

5.2 Issuing or drafting audit opinions based solely on AI analysis without application of professional judgment.

5.3 Submitting client confidential data to AI systems not approved under Section 6.1.

6. SAFEGUARDS

6.1 *Approved Vendors.* AI tools must be approved by the National Office Technology Committee. Approved vendors must maintain SOC 2 Type II certification and execute firm-standard Data Processing Agreements.

6.2 *Professional Review.* All AI-assisted work product must be reviewed by a manager or partner before inclusion in workpapers or delivery to clients. The reviewer must document their assessment of the AI output's accuracy and appropriateness.

6.3 *Workpaper Documentation.* Audit documentation must identify procedures that used AI tools, describe the tool and methodology, and explain how professional judgment was applied to AI outputs.

7. TRAINING REQUIREMENTS

All audit professionals must complete AI Fundamentals training (4 hours) before using approved AI tools. Partners and senior managers must complete the Executive AI Briefing (2 hours). Annual refresher training (1 hour) is required for all personnel.

8. INCIDENT REPORTING

Personnel must immediately report to the Engagement Partner any suspected independence impairment, data breach, or client complaint involving AI tools. The Engagement Partner shall escalate to the National Office Ethics Group within 24 hours.

9. COMPLIANCE AND ENFORCEMENT

Violations of this policy may result in disciplinary action, up to and including termination. Willful violations that result in client harm or regulatory action may be referred to professional licensing bodies.

10. EXCEPTIONS

Requests for policy exceptions must be submitted in writing to the National Office Ethics Group and approved by the Chief Ethics Officer prior to implementation. Approved exceptions are valid for one engagement only.

RELATED POLICIES: Data Classification Policy (DC-2023-001); Vendor Risk Management Policy (VR-2023-004); Professional Standards Manual Ch. 12

Figure 11: Excerpt from an AI acceptable use policy for a professional services firm. The policy follows ISO 27001 documentation standards (document owner, approval authority, version control, review date) while operationalizing AICPA independence rules and SEC auditor independence requirements. Key elements include definitions, permitted and prohibited uses, safeguards, training requirements, incident reporting, and compliance enforcement.

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 (*Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023)* 2023). ECOA fair lending enforcement has traditionally applied disparate impact theory, holding lenders responsible for discriminatory effects 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 approach 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). Then conduct the dimensional calibration exercise (autonomy, entity frame, goal dynamics, persistence). Finally, perform risk assessment across bias, accuracy, security, privacy, safety, and compliance dimensions. 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 10). 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 (Figure 12).

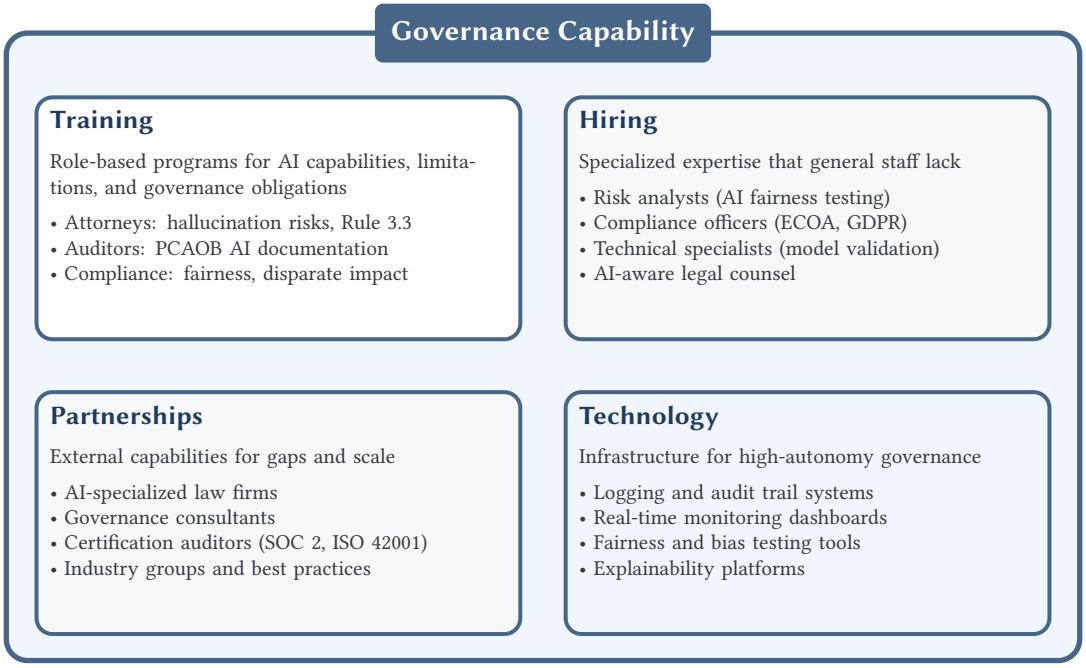


Figure 12: Four areas of sustained governance investment. Underinvestment in any area creates capability gaps that undermine the others.

Each area addresses distinct organizational needs. *Training* builds competence across existing staff—governance effectiveness depends on professionals understanding both AI limitations and their domain-specific obligations. *Hiring* fills expertise gaps that training alone cannot address; organizations serious about governance must develop or acquire specialized capabilities. *Partnerships* provide external capabilities for certification, rapidly evolving regulatory guidance, and functions that

lack economies of scale in-house. *Technology* provides infrastructure—high-autonomy systems cannot be governed with spreadsheets and manual reviews, so organizations must budget for governance tooling as part of deployment costs, not as an afterthought.

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).

Mata v. Avianca, Inc., 678 F. Supp. 3d 443 (S.D.N.Y. 2023) (2023). Sanctions order imposing $5,000 penalty on each attorney who cited six fictitious ChatGPT-generated cases. URL: https://www.courtlistener.com/docket/63107798/mata-v-avianca-inc/ (visited on 12/11/2025).

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).