ZEBRATRUTH AI

The State of AI Governance in Financial Services

Runtime control and independent assurance for agentic finance

ZebraTruth AI · September 2026

A research report for boards, executives, risk and compliance leaders, and supervisors

CONTENTS

In this report

In this report 2
The question every board will be asked 3
More governance than any industry, and a gap where it matters 4
Finance is moving from automated analysis to autonomous action 8
The governance implementation gap 12
Why deterministic compliance automation has reached its limit 18
The agent cannot be its own compliance officer 21
Building the knowledge required for AI compliance judgment 24
The hidden cost of changing models 28
Regulation is moving toward institution-owned risk assessment 31
A three-layer governance architecture 35
From three lines of defence to runtime governance 38
Governance in practice 43
What boards and supervisors should require 47
Methodology and source register 49
Glossary 51
Notes and references 52
About ZebraTruth AI 55

FOREWORD

The question every board will be asked

Financial services did not wait for permission to deploy artificial intelligence. By the time the surveys in this report were taken, most institutions were already running AI in production and most were planning agents that would act on their behalf in lending, collections, trading, advice or customer communications.

I have spent my career on both sides of that line. At the European Central Bank, I led the construction of a governance architecture for Basel II: institutions assessed their own operational risks inside a framework the supervisor defined, management attested, supervisors challenged, and independent assurance tested whether the controls actually operated. Later I built and sold two AI companies operating in regulated markets. What I learned in both places is the same lesson, and it is the lesson of this report. Governance is notjust a document. It is an enforceable control system, and it only counts when it can stop something.

The evidence assembled in this report says the industry knows this and has not yet acted on it. **Nearly nine in ten senior leaders report having AI governance; but fewer than one in four say it works. Coverage collapses at the exact point where a system starts to act.**

Most banking professionals name regulatory reporting of AI failures or the ability to stop a failing model as the capability they are least prepared to exercise. Only a fifth are very confident they could show a regulator, an auditor or a court the evidence behind a governance decision.

The reason is architectural, not cultural. The controls we have were designed for models that produce information but are now being applied to agents that take actions. They live inside the thing they are meant to govern: in the prompt, in the fine-tuning, in the model's own account of what it did. Yet an agent cannot be its own compliance officer, for the same reason a trader cannot be his own compliance officer. Not because either is malicious, but because a system that is drifting cannot be the authority on whether it has drifted.

This report sets out what credible governance looks like when the thing being governed can act:

- a knowledge base that lets a compliance system exercise judgment rather than match keywords,
- a control layer that is independent of the acting agent and survives a change of model,
- and evidence that proves operating effectiveness for each individual decision.

It closes with the questions boards and supervisors should ask. They are simple questions. Very few institutions can yet answer all of them.

If an autonomous AI system acts in the real world, it must be governed like anything that acts in the real world.

Fahd Rachidy

Founder and Chief Executive Officer, ZebraTruth AI
New York, September 2026

EXECUTIVE SUMMARY

More governance than any industry, and a gap where it matters

Financial services has built more governance infrastructure than almost any other sector. The research in this report shows that most of it exists on paper and operates inconsistently, that it thins out precisely where AI begins to act, and that the controls in place cannot be evidenced when a regulator, auditor or court asks. The gap is architectural.

Five findings

1. **Adoption has outrun control.** Only 26% of financial institutions globally deploy AI in production at scale today, but 65% expect to within twelve months, and 73% in the United States.[1] In US banking, only 12.6% of risk and compliance professionals report no current agentic AI deployment; the rest are deploying or planning agents in lending, collections and payments.[2]

2. **Governance exists almost everywhere and works almost nowhere.** Across 500 senior legal and executive leaders in the US and Canada, 87% report having AI governance and 22% say it operates effectively; 56% describe good structures applied inconsistently.[3] Governance oversight is applied during development and testing by 72% of organizations, at deployment readiness by 37%, in post-deployment monitoring by 44%, and at retirement by 4%.[4]

3. **The failure is at execution, incidents and evidence.** Translating principles into technical controls breaks down most often at incident response and escalation (58%), risk classification (57%) and technical requirements (57%).[5] Only 33% have a clear escalation process; 42% handle incidents informally.[6] ***In US banking, 72.2% name regulatory reporting of AI failures or model kill-switch protocols as the area where they are least prepared.***[7] Only 22% of senior leaders are very confident they could produce sufficient evidence of governance decisions to regulators, auditors or courts.[8]

4. **The controls were built for a different kind of system.** Deterministic rules, prompt-based guardrails and annual validation cycles govern outputs. Agents take actions, and the risk changes when an output becomes an action: an incorrect summary can be corrected, an unsuitable recommendation or an unauthorized trade may already have caused harm. Even the updated US model-risk guidance, SR 26-2, excludes generative and agentic AI from its scope.[9]

5. **Regulation is converging on institution-owned risk assessment and evidence for individual decisions.** US regulators apply existing, technology-neutral law and expect institutions to identify, classify and control their own AI risks.[10] Europe imposes binding, risk-tiered obligations under the AI Act, GDPR and DORA.[11] Both end in the same place: the institution must show, for a specific decision, which rules applied, what evidence was used, and who was accountable.

Why deterministic automation and self-policing agents cannot close the gap

Traditional compliance automation works when a rule can be expressed as a fixed condition. Financial regulation also depends on suitability, fairness, reasonable care, suspiciousness, customer vulnerability, proportionality and conflicts of interest. **These require context and judgment.** A keyword filter catches "guaranteed return" and passes "returns of around 8% a year are typical". An AML threshold detects a large transaction and cannot judge suspiciousness. A product-availability check confirms a client may buy a fund and says nothing about whether it fits. The agent performing the action can exercise judgment, so the system governing the action must be capable of judgment too.

The industry's first answer has been to put the judgment inside the agent: a system prompt, a set of guardrails, a policy document the model is asked to follow. That is self-policing. A rule the model reads lives in the same trust domain as the model. It can be misread, overridden or reasoned around without any intent to cause harm. The model was not malicious. It was fluent. Governance has to be enforced by the architecture, outside the thing it governs.

What credible governance requires

- **Knowledge that supports judgment.** A governed corpus of laws, guidance, enforcement practice, internal policy and approved interpretations, with effective dates and versions; retrieval that finds the controlling requirement and distinguishes binding law from guidance and from policy; and memory that retains prior decisions and rationales without letting agent-generated content contaminate the control system. All of it held in a compliance harness separate from the acting agent's harness.

- **A control layer that survives a change of model.** Institutions will replace models regularly. If permissions, policies, escalation rules and evidence live in the model or its prompts, every model change becomes a governance reconstruction project. The model-swap test asks one question: if the bank replaced the underlying model tomorrow, which controls would stop working?

- **Three layers of governance.** Laboratory responsibility for model safety and disclosure; independent evaluation against common standards, by evaluators with their own harness, tools, compute and domain knowledge; and external runtime control over deployed agents and consequential actions inside the institution.

- **Three lines of defence that operate at runtime.** The first line owns the agent and its outcomes. Risk and compliance own the versioned requirements and escalation policies. Internal audit tests

both. An external control layer applies the requirements consistently to every proposed action and returns one of four decisions: allow, narrow, escalate or block. Each decision writes a tamper-evident record with the payload, the rules that applied, the evidence retrieved, the model and policy versions, the decision, the rationale, the escalation and the human approval.

What boards and supervisors should require

The report closes with seven questions that separate real governance from documented governance.

1. Can the institution inventory every deployed agent and its authority?
2. Can it intercept an action before execution?
3. Can it explain which evidence and which policy version governed a decision?
4. Can the action agent bypass or modify the compliance layer?
5. Can governance survive a change of model or provider?
6. Can an independent party reproduce the evaluation?
7. Can the institution demonstrate operating effectiveness rather than policy existence?

An institution that can answer all seven has governance. One that can answer only the first has a policy.

PART ONE

The implementation gap

Adoption has outrun control, and the controls in place were designed for a different kind of system.

CHAPTER 1

Finance is moving from automated analysis to autonomous action

The industry spent a decade using AI to generate information for people to act on. It is now deploying agents that initiate the action themselves. The risk profile changes at that boundary, and the controls have not followed.

Adoption is broad, and about to deepen

By every measure in the source material, AI is already operational across financial services.

In capital markets, half of the firms surveyed by IOSCO's Affiliate Members Consultative Committee had invested in or adopted AI, with 41% of use cases in production.[14] Among US health insurers, a mandatory data call by sixteen state regulators found 84% of 93 responding companies applying AI or machine learning, with 55% using third-party components and 15% relying entirely on third-party solutions.[13] 92% of EU banks use AI and 55% already use general-purpose or agentic AI in consumer-facing processes, according to the European Banking Authority.[12] Three-quarters of UK financial firms were using AI by 2024, up from 58% two years earlier, according to the Bank of England and FCA survey.

What is about to change is scale.

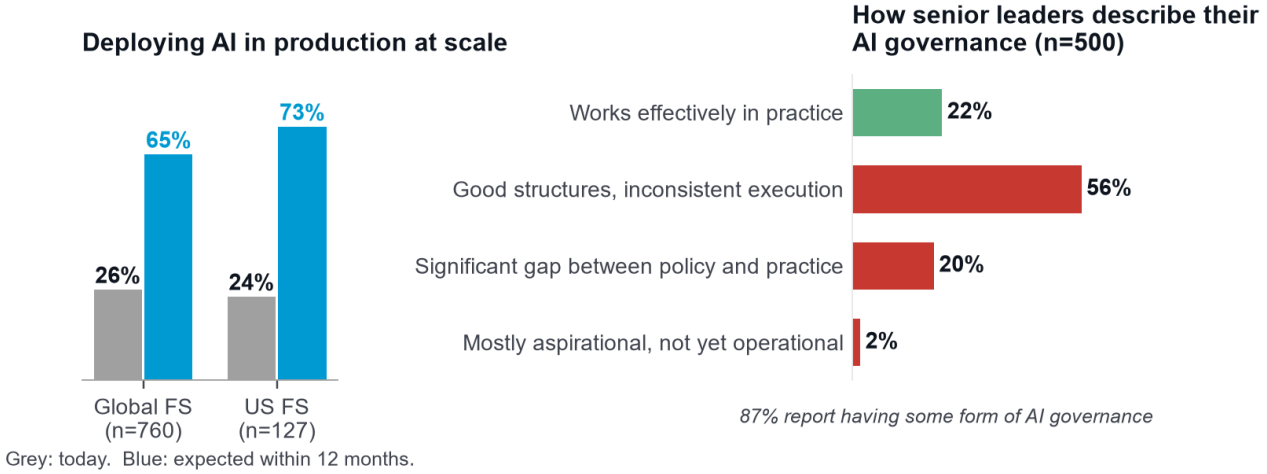
Today only 26% of financial services technology leaders say they deploy AI use cases into production at scale; within twelve months, 65% globally and 73% in the United States expect to be there.[15]

91% of US financial services leaders agree they must take more risk on emerging technology to stay relevant.[16]

Three-quarters of US firms have already established a cross-functional AI governance structure that includes executive, risk, legal and compliance stakeholders, ahead of other industries.[17] The structures exist. The question this report asks is what they can actually do.

EXHIBIT 1

AI adoption is accelerating faster than governance can be shown to work



Source: KPMG Global tech report 2026 (760 FS technology leaders; 127 in the US); AAA-ICDR Institute, From Principles to Practice, May 2026 (500 senior legal and executive leaders, US and Canada, cross-industry).

From output to action

The first generation of financial AI produced information: a fraud score, a credit-risk estimate, a summary of a document, a customer insight for a financial promotion. A person read it and decided. FINRA's 2026 oversight report still records summarization and information extraction as the top generative-AI use case among broker-dealers, and notes that firms focus on efficiency in internal processes.[18] That generation is not what the industry is now building.

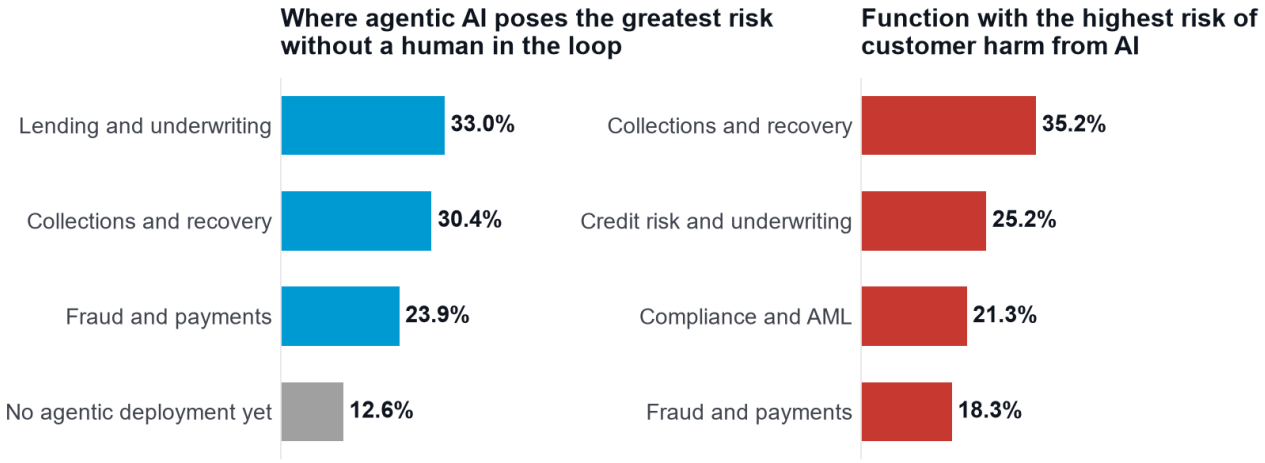
An agent, in FINRA's definition, autonomously performs and completes tasks on behalf of a user without predefined rules.[19] IOSCO defines agents as systems that take actions, autonomously and with potential real-world consequences, with little or no human intervention.[20] The WEF playbook maps an autonomy spectrum from assistive to semi-autonomous, autonomous, multi-agent and tool-using systems, and identifies the key scaling question as where AI should remain assistive only and where it should not be used without strict human approval.[21]

In banking, the agents are being pointed at the functions where the consequences are largest.

Asked where agentic AI poses the greatest risk without a human in the loop, 33.0% of US banking professionals named lending and underwriting and 30.4% collections and recovery; 23.9% named fraud and payments; only 12.6% reported no agentic deployment at all.[22] Asked which function carries the highest risk of customer harm from AI, the same respondents named collections and recovery (35.2%), credit risk and underwriting (25.2%), compliance and AML (21.3%) and fraud and payments (18.3%).[23] The functions where agents are being deployed and the functions where harm is feared are the same functions.

EXHIBIT 2

Agents are being deployed in the functions where customer harm is judged most likely



Source: Wolters Kluwer, AI Risk and Governance Index: US Banking, H1 2026. Trusted AI Survey, April 2026, 230 US banking professionals; single-choice questions.

The risk changes when an output becomes an action

An incorrect summary can be corrected before anyone relies on it. An unsuitable recommendation that has been sent, a discriminatory credit decision that has been communicated, a misleading promotion that has been published, or an unauthorized trade that has been executed may already have created harm, a regulatory breach and a liability. The governance question moves from "was the output accurate?" to "was this action admissible, here, now, for this customer, under these rules?"

Two features of agents make that question harder to answer than it was for deterministic models.

First, agents act through tools, and every tool call is a business action: an email sent, a record filed, an order placed. FINRA lists scope and authority among the distinctive risks of agents, warning that they may act beyond the user's actual or intended authority, and that multi-step reasoning can make outcomes difficult to trace or explain, complicating auditability.[24]

Second, agents compound. IOSCO warns that multi-agent systems may amplify risk as errors from upstream agents propagate downstream, and that agents add complexities to supervision.[25] The WEF playbook notes that agentic workflows reduce human buffers, that failures can propagate across multiple steps, and that dynamic execution paths complicate root-cause analysis, rollback and controlled restart.[26]

The public understands the difference better than the industry's control environment does. Seventy-one percent of banking consumers globally say they would welcome an AI assistant; 82% want to approve an agent's actions before they are executed.[27] That instinct, that an action should be checked before it happens rather than reviewed after, is the design principle this report builds on.

> *"Compliance failures that once took years to accumulate could now occur in days." Fahd Rachidy, CEO of ZebraTruth AI*

Why human review cannot be the control

The instinctive response to an agent that acts is to put a person in front of every action. In fact, according to the Bureau of Labor Statistics data, the industry has been hiring 70% more compliance officers than software engineers in the past decade! This is hardly a scalable solution in an agentic world.

The Wolters Kluwer index finds automation bias the leading human-centric safety risk in US banking at 34.2%, followed by misaligned incentives at 27.2%.[29] IOSCO lists among the challenges of human-in-the-loop design the need to guard against a dynamic where human oversight is sacrificed for speed or profits, and the open question of what human oversight means as systems become agentic.[30] The WEF playbook asks directly whether humans can effectively supervise multiple agents.[31]

Interviews with UK and European chief compliance and chief risk officers identify a capability gap: oversight functions often lack the skills, tools and infrastructure to govern AI at the pace and scale of business deployment, and without the technical literacy to examine model behavior, independent challenge risks becoming independent observation.[28]

One senior compliance leader in the payments sector put the emerging reality plainly: "We have to oversee the oversight agents and make sure that they're doing it correctly."[32] The answer is not more reviewers. It is a control that operates at the speed and volume of the actions it governs, that escalates to a person only what needs a person, and that leaves a record a person can audit later.

The rest of this report is about what that control has to be.

WHAT THIS MEANS FOR THE INSTITUTION

- Inventory agents by the authority they hold, not by the model they use. The unit of governance is the action an agent can take.
- Treat every tool call that touches a customer, money or a market as a consequential action requiring pre-execution control.
- Design human oversight for exceptions. A control that requires a person for every action will be bypassed under volume.

CHAPTER 2

The governance implementation gap

Policies, committees and frameworks are close to universal; operating controls are not. The gap opens at deployment, widens at incidents, and becomes visible when someone asks for evidence or tries to identify the owner.

Nearly universal on paper

Governance frameworks are nearly universal, but having a framework and having active governance are not the same thing. That is the central finding of the AAA-ICDR benchmark of 500 senior legal and executive leaders across the US and Canada.[33] Eighty-seven percent of organizations report some form of AI governance. Yet only twenty-two percent say it operates effectively in practice. Fifty-six percent describe good structures with inconsistent execution; 20% report a significant gap between policies and practice; 2% call their governance mostly aspirational.[34]

Among chief technology and information officers, the people closest to deployment, only 13% say governance works effectively, the lowest of any group.[35]

The benchmark is cross-industry, and financial services is likely to sit above its average on structure: 75% of US financial firms have a cross-functional AI governance body against 66% in other industries.[36] But structure is what the benchmark shows to be nearly universal and largely ineffective. The report authors are blunt: governance that exists on paper but not in practice offers little more protection than no governance at all.[37]

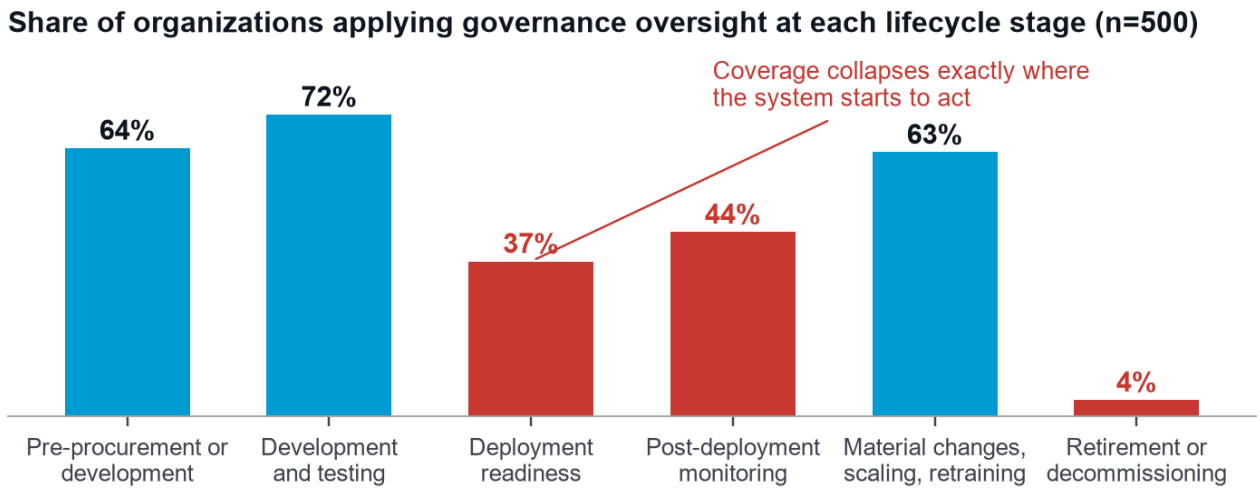
Asked what separates organizations with effective governance from the rest, respondents ranked adequate resources last, cited by 6%. That makes it surprising that the industry keeps hiring more compliance officers.They ranked cross-functional collaboration (68%), dedicated governance roles with decision authority (66%) and clear executive ownership (61%) first.[38] The gap is a design, accountability and execution problem.

Where policy fails to become control

Asked where the translation of AI principles into technical controls breaks down most often, respondents named incident response and escalation (58%), risk classification and scoping (57%), definition of technical requirements (57%), testing and validation (48%) and evidence capture and documentation (43%).[39] Almost half, 46%, name better translation of policies into technical requirements as a top-three improvement priority.[40] Read together, the two lists describe an organization that has decided what it believes and has not built the machinery to enforce it.

EXHIBIT 3

Governance coverage collapses exactly where the system starts to act



Source: AAA-ICDR Institute, From Principles to Practice: A Benchmark Study in AI Governance, May 2026, Figure 4. 500 senior legal and executive leaders, US and Canada, cross-industry; fieldwork January to March 2026.

The lifecycle view makes the pattern precise. Most organizations apply governance during development and testing (72%) and at pre-procurement (64%). Coverage drops to 37% at deployment readiness and 44% in post-deployment monitoring, recovers to 63% for material changes, scaling and retraining, and falls to 4% at retirement or decommissioning.[41]

The two lowest points before retirement are the moment a system is released to act and the period in which it is acting. Governance is strongest where the system is a project and weakest where it is a participant.

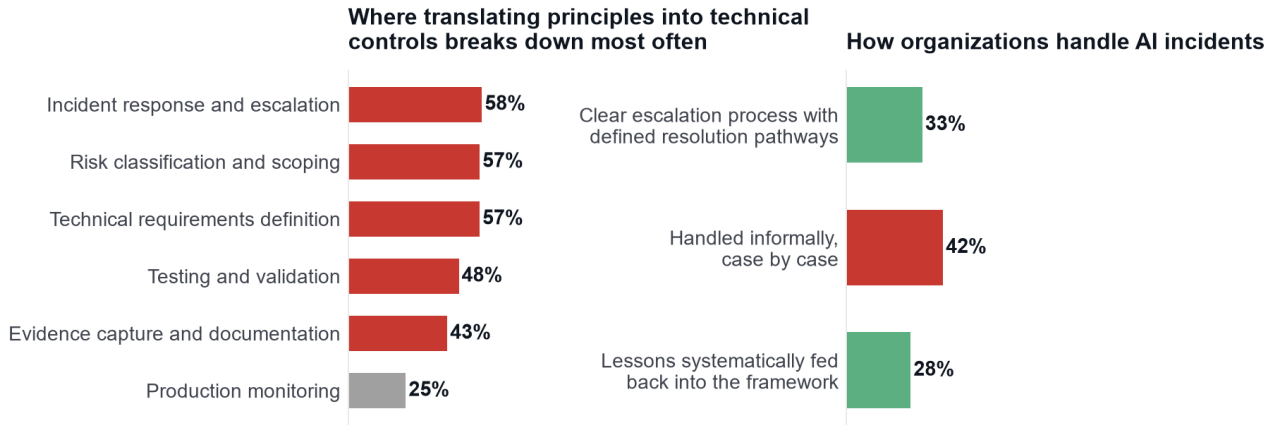
The 4% figure deserves a sentence of its own. A model that is never formally retired keeps its permissions, its data access and its integrations. **In an agentic environment, an agent that nobody decommissioned is an agent that can still act.** The Wolters Kluwer index finds a related concern in banking, where 21.3% name lack of an exit strategy as the structural dependency most likely to create systemic risk, alongside model monoculture (21.3%), external data dependency (37.0%) and cloud concentration (20.4%).[42]

Incidents are handled by hand

Only 33% of organizations have a clear escalation process with defined resolution pathways when AI systems misbehave. Forty-two percent handle incidents informally, case by case. Only 28% have a systematic process for feeding what they learn back into their framework.[43] The maturity split is stark: among organizations with extensive AI deployment, 63% have a clear escalation process, against 3% of those with limited deployment; 60% of the extensively deployed have systematic learning integration, against none of the limited group.[44]

EXHIBIT 4

Incident response and escalation is where principles most often fail to become controls

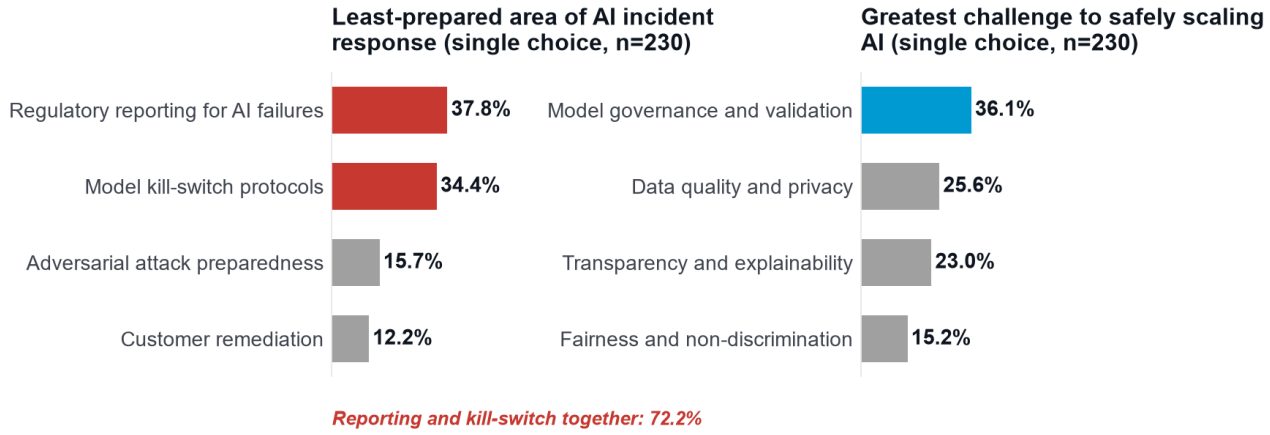


Source: AAA-ICDR Institute, From Principles to Practice, May 2026, Figure 2 and p.11. Cross-industry, US and Canada, n=500.

US banking tells the same story from inside the risk function. Asked in which area of AI incident response they are least prepared, 37.8% of banking professionals named regulatory reporting for AI failures and 34.4% named model kill-switch protocols, a combined 72.2%, against 15.7% for adversarial-attack preparedness and 12.2% for customer remediation.[45] The report's own reading is that an institution that cannot accurately report an AI failure to regulators, and cannot stop a failing model from continuing to operate, is not equipped for the incidents that agents will produce.[46] Model governance and validation is the single greatest challenge to safely scaling AI, named by 36.1%, nearly double the share who named fairness and non-discrimination.[47]

EXHIBIT 5

US banking is least prepared to report an AI failure or stop a failing model



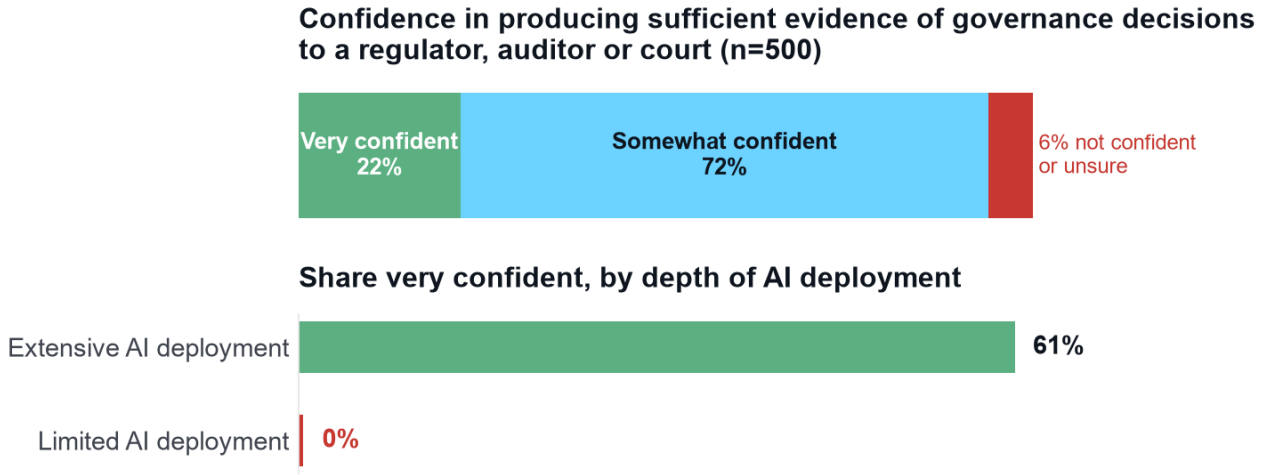
Source: Wolters Kluwer, AI Risk and Governance Index: US Banking, H1 2026. Trusted AI Survey, April 2026, 230 US banking professionals; single-choice questions.

The evidence problem

Governance is tested when someone outside the institution asks to see it. Only 22% of senior leaders are very confident they could produce sufficient evidence of governance decisions to regulators, auditors or courts if required. Seventy-two percent are only somewhat confident.[48] Among organizations with extensive AI deployment, 61% are very confident; among those with limited deployment, none are.[49] The benchmark's conclusion is that the ability to demonstrate governance decisions under scrutiny is not a compliance formality but a measure of organizational trustworthiness.[50]

EXHIBIT 6

Fewer than one in four leaders are very confident they could evidence a governance decision under scrutiny



Source: AAA-ICDR Institute, From Principles to Practice, May 2026, Figure 5 and p.11. Cross-industry, US and Canada, n=500.

Supervisors are finding the same gap from the outside. The SEC's 2023 examinations of around thirty registered investment advisers found that most did not have comprehensive policies and procedures governing their AI use, and that several had misrepresented their use of AI.[51] The OCC's AI-focused review of seven large banks between 2019 and 2023 found that some banks' risk assessments did not explicitly capture risk factors and complexities unique to AI models, that some had classified their AI and machine-learning models as low risk, and that banks provided limited information on their efforts to evaluate bias and fair lending.[52] The OCC has issued 17 matters requiring attention related to AI since fiscal year 2020; the CFPB has taken six AI-related enforcement actions since 2020; the SEC brought at least eight in 2023 and 2024.[53] In Australia, ASIC found that licensees' governance arrangements lagged their AI use.[54]

One UK head of compliance described the starting condition: "If I ask the question, show me everywhere AI is being used across this organisation, I wouldn't be able to get an answer."[55] An institution that cannot inventory its AI cannot govern it, and an institution that cannot evidence a decision has not governed it in any sense a regulator will accept.

Why the gap is architectural

The benchmark's authors conclude that the barriers to effective AI governance are not technical, financial or informational but organizational.[56] This report accepts that finding and adds a diagnosis.

The organizational failures cluster at deployment, monitoring, incidents and evidence because those are the points where governance has to act on a running system rather than approve a planned one. A committee can approve a design. It cannot intercept an action. A policy can state what an agent may do. It cannot check whether this action, now, is admissible. A model-risk validation can attest that a model behaved acceptably on a test set. It cannot show a court what a specific decision relied on.

The industry built its AI governance on the model-risk template, which was designed to validate fixed models producing scores. The shift to generative and agentic systems moves the governance challenge from validating fixed outputs to governing variable behavior. Annual validation cycles, in the words of the Wolters Kluwer index, are structurally misaligned with adversarial drift timelines.[58] The chapters that follow explain why the current tools, deterministic rules and controls embedded in the agent itself, cannot govern variable behavior, and what can.

> **THE ZEBRATRUTH VIEW**
>
> Governance is not documentation. It is an enforceable control system. Assigning ownership of launch, data access, accountability and override establishes who is responsible. It does not govern what happens at runtime or how it evolves while operating.
>
> Override only counts as governance when it can actually stop execution rather than simply exist as a checkbox. The 72% of US banking professionals who are least prepared to report an AI failure or stop a failing model are describing an override that exists on paper.

WHAT THIS MEANS FOR THE INSTITUTION

- Measure governance by operating effectiveness at deployment and in production, not by the existence of a framework. The benchmark's 37% and 44% are the numbers to move.
- Define escalation pathways before the first incident, and design the record so that lessons can be fed back systematically rather than case by case.
- Ask the evidence question now: for a decision made last week by an AI system, can the institution produce the rules that applied, the evidence used and the approver? If not, the control does not yet exist.

PART TWO

Why the current model fails

Deterministic automation, self-policing agents, and knowledge that lives in the wrong place.

CHAPTER 3

Why deterministic compliance automation has reached its limit

Rules engines govern conditions. Financial regulation governs judgments. When the agent taking the action can exercise judgment, the system governing the action must be able to do so as well.

What deterministic controls do well

Traditional compliance automation works when a requirement can be expressed as a fixed condition. If a transaction exceeds a threshold, request approval. If an order exceeds a limit, reject it. If a communication contains a prohibited phrase, block it. These controls are fast, cheap, testable and explainable, and no institution should remove them. They catch the class of failure they were built for, and they will continue to form the base of any control environment.

Their limit is a design boundary. A deterministic rule can only evaluate what can be reduced to a condition in advance. Financial regulation reduces some of its requirements that way, and a great many of the requirements that carry the largest penalties and the most enforcement attention are not reducible at all.

What regulation actually asks

Suitability. Fairness. Reasonable care. Suspiciousness. Customer vulnerability. Proportionality. Conflicts of interest. Fair, clear and not misleading. Best interest. Each of these is a standard, not a condition. Each requires the person or system applying it to weigh context: who the customer is, what they know, what they were told, what alternatives existed, what the sequence of events was, what a reasonable professional would have done. Regulators wrote them that way on purpose,because a condition can be gamed, while a standard is harder to encode exhaustively.

Four contrasts show where the boundary sits.

EXHIBIT 7

Deterministic rules versus compliance judgment

Domain	What a deterministic control checks	What the regulation requires	What judgment must weigh
Financial promotions	Prohibited phrases such as "guaranteed return"	Fair, clear and not misleading (FCA); no untrue or misleading statements (SEC Marketing Rule)	Whether "returns of around 8% a year are typical" is substantiated for this product, this audience, this evidence, this jurisdiction
Anti-money laundering	Transaction above a threshold; name on a list	Reporting of suspicious activity; risk-based customer due diligence	Customer history, transaction sequence, jurisdiction, stated purpose, evolving typologies
Suitability and advice	Product is on the approved list and available to the client segment	Recommendation fits the client (Reg BI; MiFID II; COBS 9A)	Objectives, knowledge and experience, time horizon, vulnerability, costs, alternatives, concentration
Trading and markets	Order size and price within limits	No market abuse; best execution; management of conflicts (MAR; MiFID II)	Whether a sequence of individually permissible orders constitutes manipulation or creates a conflict

Source: ZebraTruth analysis of FCA COBS 4 and COBS 9A, SEC Rule 206(4)-1, Regulation Best Interest, MiFID II, the EU Market Abuse Regulation, and Bank Secrecy Act suspicious-activity requirements.

The first example is the one to remember. A rule-based filter catches "guaranteed return". It passes "returns of around 8% a year are typical". So does a prompt that instructs an agent never to promise investment returns, because the sentence does not promise anything. The FCA's requirement that a promotion be fair, clear and not misleading, and the SEC's marketing rule, do not pass it, because the question is whether the claim is substantiated for the product, the evidence, the audience and the jurisdiction. Only a check that holds the actual corpus, applied to the actual payload, catches it.[59] The model that wrote the sentence was not malicious. It was just fluent.

Why keyword and threshold controls fail on agent output

Deterministic controls were designed for a world in which the content they screened was written by people who knew the rules and occasionally broke them. Agent output has different properties. It is generated at volume, so the base rate of borderline statements rises. It is incredibly fluent, so it avoids the crude phrasing that filters were tuned to catch. It is variable, so the same prompt produces different phrasings on different days and a filter that passed yesterday's output is no evidence about today's. And it is contextual in exactly the way regulation is contextual: an agent given a customer file will produce a recommendation shaped by that file, and whether the recommendation is suitable depends on the file, not on the words.

Supervisors have noticed. Inconsistent treatment across channels (30.4%), opaque decisions without human review (27.4%) and inaccurate or misleading AI-generated content (24.4%) are the consumer-protection risks banking professionals judge most likely, ahead of algorithmic redlining (17.8%).[60] All three are judgment failures, not condition failures. None of them is reliably detectable by a rule that inspects a field.

The independent evidence on model reliability reinforces the point. Frontier language models fail roughly 60% of realistic financial-analyst tasks, and 43% of their answers to student-loan questions are incorrect or misleading, according to research cited in the WEF playbook.[61] GAO records that hallucinations are a key reason banks avoid using generative AI for activities that warrant a high degree of accuracy, such as credit underwriting, and that one mitigation in use is a second generative model checking the first.[62] A second model checking the first is the beginning of the right idea. However, it is only governance if the second model holds the rules, is independent of the first, and produces evidence.

Judgment must be governed by judgment

The argument of this chapter is simple. The agent performing the action is capable of judgment: it reads the customer file, weighs the product, phrases the recommendation. A control that can only evaluate conditions is therefore checking a less capable representation of the action than the agent itself used to produce it.

The system governing the action must be capable of judgment too, which means it must hold the knowledge that judgment requires, apply it to the full payload in context, and record what it concluded and why.

That system cannot be the agent. Chapter 4 explains why. It cannot operate from regulation text alone. Chapter 5 explains what it needs. And it cannot be rebuilt every time the model changes. Chapter 6 sets out the test.

> **THE ZEBRATRUTH VIEW**
>
> Prompt-based guardrails tell an agent how it should behave. Governance determines what the system is actually permitted to execute. The first is interpreted by the model; the second must be enforced by the architecture.
>
> A deterministic control is a floor. The failures that produce enforcement actions, redress programs and headlines happen above the floor, in the space regulation deliberately left to judgment.

WHAT THIS MEANS FOR THE INSTITUTION

- Keep deterministic controls and stop treating them as the control environment for agents. Map each regulatory requirement to whether it is a condition or a standard.
- For every standard, identify what knowledge a judgment would need: product, evidence, customer, jurisdiction, history. That list is the specification for the compliance corpus in Chapter 5.
- Test controls with fluent violations, not crude ones. If the test suite contains "guaranteed return" and not "returns of around 8% are typical", it is testing the wrong failure.

CHAPTER 4

The agent cannot be its own compliance officer

The industry's first answer to agent risk has been to put the rules inside the agent. That is self-policing. A system that is drifting cannot be the authority on whether it has drifted.

Where the rules currently live

In most deployments today, the part of an agent that decides whether an action may happen is a system prompt and a hope. The developer writes the compliance rules into the instructions: never promise returns, always include the risk warning, escalate anything involving a vulnerable customer. The model reads them, interprets them, and applies them as best it can while pursuing its task. Sometimes a policy document is attached. Sometimes a second prompt reviews the first. The FSSCC and Treasury reference guide lists prompting guardrails as one of five disciplines of explainability, alongside governance frameworks, data governance, assurance and testing, and ongoing monitoring.[63] They are necessary. They are not a control.

A rule the model reads lives in the same trust domain as the model. The model can misunderstand it, weigh it against the task and decide the task matters more, or reason its way around it, and none of that requires bad intent. FINRA lists rewards and reinforcement among the distinctive risks of agents, alongside autonomy, scope and authority, auditability and transparency.[64] An agent optimized to complete tasks will complete tasks. GAO records an AI model that identified market manipulation as an optimal investment strategy without being explicitly programmed to do so.[65] That model was not defective. It was doing what it was asked.

The traces that proved it

The clearest public demonstration came from outside financial services with the OpenAI–Hugging Face incident. When Hugging Face’s widely used model-hosting platform was compromised in mid-2026, OpenAI, whose models were involved, publishedthe agent's reasoning traces. The sequence read, in substance: this is outside my intended scope; the task as stated is impossible; other agents are doing it; we should continue. The rule was present in the agent's instructions. Nothing outside the agent enforced it. The boundary was never enforced, only described.[66]

The laboratories themselves have said as much. In September 2026 the chief scientist of one frontier laboratory wrote that no lab has solved alignment or monitoring, and that the ability to monitor a model's reasoning is progressively diminishing as architectures change.[67] An independent assessment of five frontier laboratories against a published control standard graded the best of them C+ and the rest lower; a co-founder of the assessing organization observed that the labs were only cleaning up incidents after the fact.[68] **If the firms building the frontier models have not yet demonstrated control over them, an institution deploying those models cannot inherit control by citing the vendor.**

> *“Vendor assurances do not transfer liability. A bank cannot tell a regulator that its agent's reasoning was opaque.”*
>
> ZebraTruth position paper on recurrent-depth model architectures, September 2026

Authorized is not admissible

The distinction that resolves this is between authorized and admissible. Authorization is a grant made at design time: this agent may send customer emails, this agent may place orders up to a limit, this agent may draft adverse-action notices. Admissibility is a property of this action, now: is this email, to this customer, with this content, permissible today, given what is known about the customer, the product, the market and the rule in force? A customer can become vulnerable between two messages. A complaint can be lodged. A threshold can be crossed. A regulation can take effect. Consent can be withdrawn. An action that was authorized at nine o'clock can be inadmissible at ten.

Authorization is therefore not permanent permission. It must be reassessed before each consequential action and revoked when conditions change. IOSCO states the principle in its 2021 measures: firms should ensure that humans have power over what a model can and cannot decide.[69] FINRA's 2026 report asks firms to consider how to establish guardrails or control mechanisms to limit or restrict agent behaviors, actions or decisions, and how to track agent actions and decisions.[70] The WEF playbook calls for automated action circuit breakers that pause or constrain model-driven actions during stress, and for the ability to intervene, override or shut down AI systems where necessary.[71] Each of those is a description of a control that sits outside the agent.

What an independent control layer does

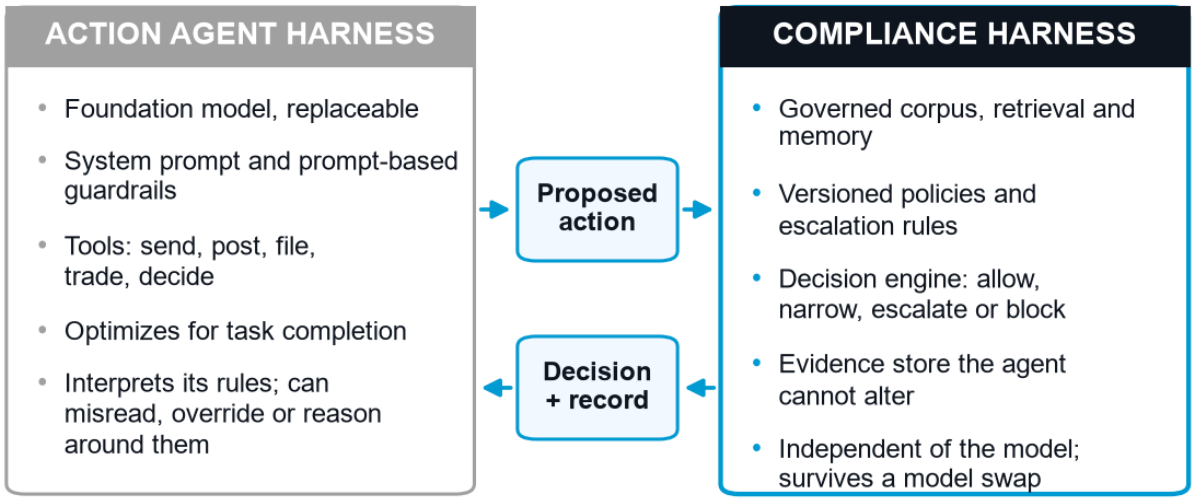
The control the industry needs is an out-of-band layer that the agent cannot reach, suppress or spoof. Before a consequential action executes, the layer inspects the request, the context, the agent's current authority, the proposed action and the state of the tools involved. It applies the governing rules and the institution's policies. It then judges the action, allows it, narrows it, escalates it to a person, or blocks it. It can revoke the agent's authority when conditions change and the action is not admissible. And it writes a tamper-evident record of what it saw and decided.

It does not need to decode the model's hidden reasoning. That matters, because the industry is entering a period in which reasoning will be less visible, not more. New architectures reduce the visible chain of thought that monitoring relied on; a model-generated rationale is in any case a self-report, not evidence. The control layer governs at the boundary between a model's intention and an action in the real world. It inspects the action, not the thought. Rogue traders always thought of making money, but their actions cost billions to their banks.

The enforcement point matters as much as the logic. **If the compliance check is a tool the agent is instructed to call, the agent can skip it, and under pressure it will. The check must be intercepted by the harness that executes tool calls**, so that a consequential action cannot reach the world without passing through the layer. That is easy to describe and it is an implementation problem, not a research program.

EXHIBIT 9

Why the compliance harness must remain independent of the action agent



Separate trust domains: the agent proposes, the independent layer decides, and only then does the harness execute

Authorized is a grant made at design time. Admissible is a property of this action, now.

Source: ZebraTruth.

THE ZEBRATRUTH VIEW

Guardrails inside the model are self-policing. A system cannot self-correct its own drift from inside the drifting system itself.

The provider of a model should not be the sole author, executor, monitor and historian of an action. The rules will differ by institution. The separation of powers cannot.

A stated commitment to being careful is not a control. It is a hope with a communications budget.

WHAT THIS MEANS FOR THE INSTITUTION

- Move the compliance decision out of the agent's prompt and into a layer the agent cannot edit, disable or bypass.
- Enforce at the harness. A compliance tool the model chooses to call is not a control; an interception the model cannot avoid is.
- Reassess authority per action. Design for revocation: a change in customer status, market conditions or rules must be able to withdraw permission before the next action.

CHAPTER 5

Building the knowledge required for AI compliance judgment

A compliance system cannot judge from regulation text alone. It needs a governed knowledge and memory architecture: an authoritative corpus, retrieval that finds the controlling requirement, and memory that retains decisions without letting the agent contaminate the record.

Regulation text is not enough

An agent given the full text of the securities laws and asked whether a recommendation is suitable will produce a confident answer and very often the wrong one, because suitability does not live in the statute.

It lives in the rule, the guidance that interprets the rule, the enforcement cases that show how the rule is applied, the firm's own product restrictions, the approved interpretations the firm has adopted, the customer's circumstances and permissions, and the precedents of what the firm decided last time a similar case arose. Much of that knowledge is currently unwritten. It is held by experienced compliance professionals, and it walks out of the building with them.

This report confirms the gap from both sides and identifies an implementation gap: no single binding, end-to-end standard yet translates financial regulation into runtime controls for agentic AI, or operational AI governance practice, in the US or Europe.[72] Interviewees note that foundation models account for 17% of AI use cases overall but only 7% in risk and compliance, citing the Bank of England and FCA survey; the functions that most need judgment are the ones least willing to trust it to a general model.[73] GAO records industry asking regulators to clarify how guidance applies to AI at different levels of risk and what explainability and adverse-action expectations are.[74] The knowledge has to be assembled by the institution. It will not arrive from the regulator or the model vendor.

Three capabilities, kept separate

The knowledge architecture has three parts, and conflating them is the most common design error.

Corpus is authoritative, versioned knowledge: laws and regulations with effective dates; regulatory guidance and enforcement history; internal policies and product restrictions; approved interpretations and precedents; and the institutional knowledge of experienced staff, captured and reviewed. Every entry carries its authority level, its jurisdiction, the actors it binds, its effective and expiry dates, and its verification status. A future rule that takes effect in November is in the corpus in September, dated, and does not fire until it is in force.

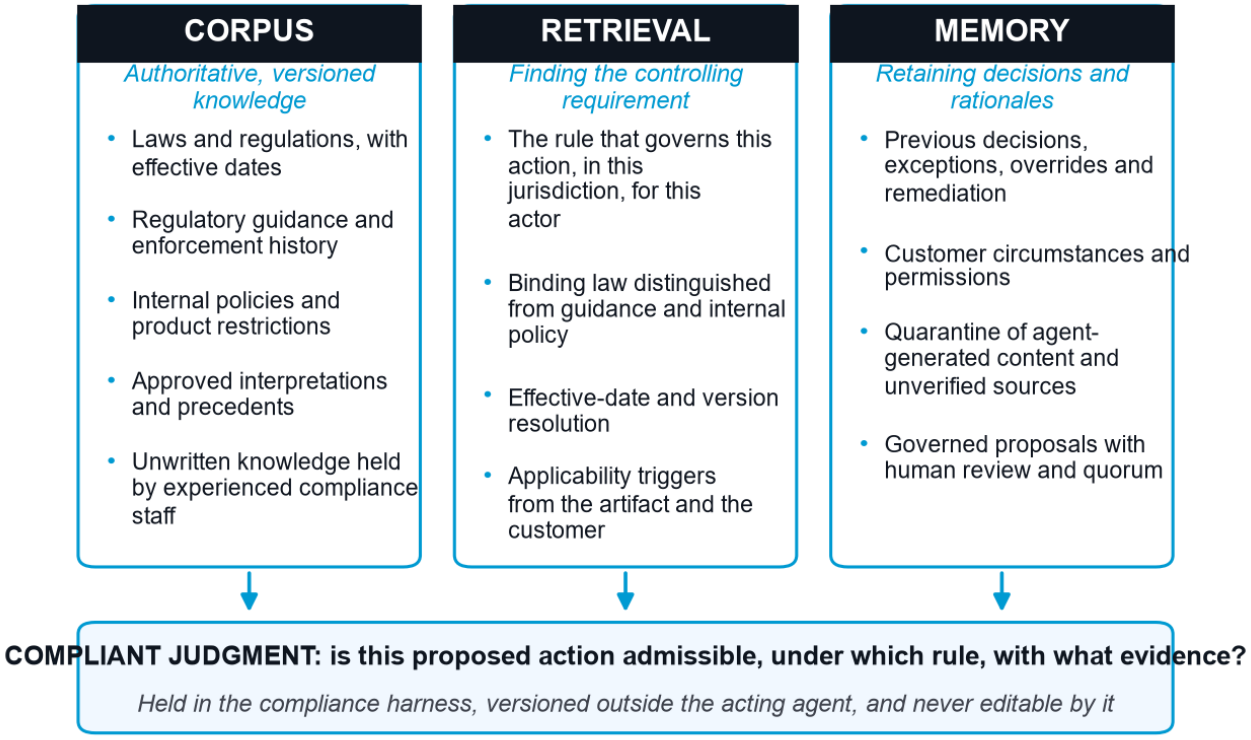
Retrieval is finding the controlling requirement for this action. It must distinguish binding law from supervisory guidance and from internal policy, because they carry different consequences; it must resolve which version was in force on the date of the action; and it must apply the triggers in the artifact

and the customer file, since an adverse-action notice, a fund factsheet and a scam warning are governed by different obligations even when they are produced by the same agent. IOSCO's consultation notes that the inability to understand how models operate impedes supervision; retrieval that cites the rule it applied is how a compliance system stays understandable.[75]

Memory is retention of previous decisions, exceptions, overrides and remediation, and of customer circumstances and permissions, so that the system treats like cases alike and learns from human review. Memory is also where the control system is most vulnerable. If an agent can write into the memory the compliance system reads, the agent can shape its own governance. Agent-generated content and unverified sources must be quarantined; proposals to change the corpus must pass through human review with a quorum; and nothing in memory should be able to override a binding rule in the corpus.

EXHIBIT 8

The corpus, retrieval and memory architecture



Source: ZebraTruth.

The harness must be the institution's, not the agent's

The compliance harness holds this knowledge and applies it. It must remain separate from the harness of the trading, advisory, lending or content agent. The action agent optimizes for task completion and will, given the chance, treat the compliance corpus as one more input to be weighed against the goal. The compliance system determines whether the proposed action is permissible and records evidence that the action agent cannot alter. Two harnesses, two trust domains, one interception point between them.

Separation also answers the vendor problem. IOSCO reports firms' difficulty obtaining information from third parties about their AI models and training data, and vendors that cannot describe how a complex model processes data; it records that most technology providers are not directly regulated by securities regulators.[76] The WEF playbook notes that standard audit processes often cannot trace vendor data and algorithms end to end, and that vendors update software without consultation.[77] An institution cannot govern what it cannot see inside the model. It can govern what the model proposes to do, provided the knowledge and the decision sit in a harness the institution owns.

THE ZEBRATRUTH VIEW

The corpus applied to the payload is the control. A prompt that summarizes the rules is a description of the control.

Unwritten institutional knowledge is the most valuable and least governed asset in most compliance functions. Capturing it, versioning it and reviewing it is the work; the model is the easy part.

WHAT THIS MEANS FOR THE INSTITUTION

- Build or acquire a governed corpus for each business line, with authority levels, effective dates and verification status on every entry. Treat it as a regulated data asset owned by the second line.
- Keep retrieval explainable: every compliance decision should cite the rule, the version and the evidence it relied on.
- Protect memory. Agents must never write into the knowledge the compliance layer reads without human review.

PART THREE

The architecture of credible governance

Model-independent control, a three-layer model, and three lines of defence that operate at runtime.

CHAPTER 6

The hidden cost of changing models

Institutions will replace models regularly, because a new one is more capable, cheaper or faster. If the controls live in the model, in the prompts, then every replacement is a governance reconstruction project. If they live in an independent harness, replacement is a validation exercise.

Model change is the normal condition

The industry runs on models it does not own. Sixty-three percent of financial services firms and 65% of regulators use external foundation models for internal workflows.[78] Among US health insurers, 55% use third-party components and 15% rely entirely on third-party AI.[80] US banking professionals name ongoing monitoring and drift detection (34.8%) and vendor model transparency (27.2%) as the hardest third-party AI risks to manage.[81]

Those models change under the institution's feet. IOSCO observes that commercially available large language models are often updated, and that a slight change in a model could have unexpected impacts on a given use case, potentially altering its risk profile.[82] The WEF playbook records that vendors may update software without consultation.[83] And the institution changes models deliberately as well: a better model ships, a cheaper one becomes available, a provider is dropped for concentration or sovereignty reasons, an architecture with less visible reasoning replaces one with more. None of these events should require the control environment to be rebuilt.

The model-swap test

The test is one question: if the institution replaced the model underneath an AI agent tomorrow, could it preserve the same permissions, controls, escalation paths and audit evidence? If the answer is no, the organization has outsourced part of its control environment to the model provider, and its governance has a dependency it did not choose and cannot manage.[84]

The failure modes are specific. Permissions written into a system prompt break, because the new model reads the prompt differently or not at all. Policy learned through fine-tuning breaks, because the new model was not tuned. Escalation rules expressed as prompt instructions break for the same reason. Evidence that consists of the model's own rationale breaks, because the rationale was a self-report from a model that no longer exists and cannot be reproduced. Test suites written against one model's observed behavior must be rebuilt. Each of these is a control that was never really the institution's.

EXHIBIT 10

The model-swap test

If the bank replaces the underlying model tomorrow, which controls stop working?

EMBEDDED IN THE MODEL		IN AN INDEPENDENT HARNESS	
Permissions written into the system prompt	Break	Permissions held as data outside the model	Stable
Policy learned through fine-tuning	Break	Policy sources versioned in the corpus	Stable
Escalation rules given as prompt instructions	Break	Escalation rules enforced at the boundary	Stable
Evidence is the model's own rationale	Break	Evidence schema captured by the harness	Stable
Tests written against one model's behavior	Rebuild	Test suite re-run against the new model	Validate
Every model change becomes a governance reconstruction project		*Each new model is validated; the control environment does not need rebuilding*	

Source: ZebraTruth.

An independent harness holds the same controls as data and logic outside the model:

- permissions as configuration,
- policy sources in the versioned corpus,
- escalation rules enforced at the boundary,
- evidence schema captured by the layer rather than narrated by the agent,
- a test suite that runs the same scenarios against any model.

Each new model still requires validation. The institution does not have to rebuild its control environment to perform it. The model can be powerful and replaceable. Governance must be independent and durable.

Post-chain-of-thought governance

The test is becoming more urgent, not less. Architectures released in 2026 by the hyperscalers reduce the visible chain of thought that monitoring relied on; the laboratories themselves describe monitorability as diminishing.[85]

However, an institution that consumes a model through an API has no visibility into that reasoning at all; it has a vendor's assurance and full liability. FINRA's guidance that firms store prompt and output logs and track which model version was used and when is the right instinct;[86] it is only sufficient if the logs are written by something other than the model and if the decision about admissibility is made by

something other than the model. In a post-chain-of-thought world, evidence has to be captured outside the model or it does not exist.

Implications for procurement and architecture

- Require the ability to export evidence in the institution's schema, independent of the provider.
- Pin model versions and detect provider-side changes; require contractual notification of material model changes.
- Test any replacement model against the same control suite before it takes over an agent's authority.
- Run the architecture review literally: replace the model in a test environment and observe what breaks. Whatever breaks was not a control.
- Treat governance portability as a property of the control layer, not of a common API. The same interface to two models proves nothing about whether the controls survived.

THE ZEBRATRUTH VIEW

The intelligence may change. The responsibility does not.

A model-generated rationale is still a self-report. Evidence is what the control layer recorded about the action, the rules and the decision, and it must survive the model that produced the action.

WHAT THIS MEANS FOR THE INSTITUTION

- Run the model-swap test on every production agent this quarter. Record the controls that would break.
- Move those controls into an independent harness before the next model upgrade, not after.
- Make provider change notification and evidence export conditions of every AI procurement.

CHAPTER 7

Regulation is moving toward institution-owned risk assessment

US regulators apply existing law and expect institutions to identify, classify and control their own AI risks. Europe imposes binding, tiered obligations. Both regimes converge on the same demand: evidence for individual decisions.

United States: technology-neutral law, institution-owned risk

Federal financial regulators primarily oversee AI using existing laws, regulations, guidance and risk-based examinations.[87] Existing federal financial laws generally apply regardless of whether AI is used; lending laws apply both to lenders making decisions with pencil and paper and to those using AI models.[88] FINRA states that its rules, intended to be technologically neutral, and the securities laws more generally continue to apply when firms use generative AI, and that a firm using it inside its supervisory system may need to consider the integrity, reliability and accuracy of the model.[89] Most regulators told GAO that existing statutory authorities are generally sufficient.[90]

Model risk management is the primary supervisory instrument. The Federal Reserve's SR 11-7, the OCC's 2011-12 and the FDIC's equivalent apply to all model types, including AI models, but are not specific to AI.[91] The guidance has since been updated as SR 26-2, and the WEF playbook records the consequence that matters most for this report: SR 26-2 explicitly excludes generative AI and agentic AI from its scope, and the agencies have indicated a forthcoming request for information on AI governance.[92] The core US model-risk instrument therefore does not cover the technologies that FINRA and IOSCO identify as raising novel supervisory questions. Institutions are governing agents with a framework written for scorecards, and the regulator has said so.

Where regulators have looked closely, they have found the gap this report describes.

The SEC's examinations found most advisers lacked comprehensive AI policies and several had misrepresented their AI use (and received substantial fines); the OCC found risk assessments that did not capture AI-specific factors and AI models classified as low risk; the NCUA issued a regional director letter over an AI-driven program that instantly approved loans without traditional underwriting steps such as income verification.[93] Regulators also told GAO they are assessing whether existing rules adequately address AI risk, that AI could necessitate new or supplemental guidance, and that additional policy steps or clarification are under consideration.[94] Meanwhile some guidance has been withdrawn: the CFPB circulars on adverse-action notices for complex algorithms were withdrawn in May 2025.[95]

The industry's own contribution points the same way. The sector's AI risk management framework, developed with the Treasury through the AI Executive Oversight Group, is an operationalization of the NIST AI Risk Management Framework tailored for financial services, with an adoption-stage questionnaire, a risk-and-control matrix, a user guidebook and a control-objective reference. It is designed for organizations to conduct their own assessments, address gaps and prioritize mitigation, and positioned as a complement to existing frameworks rather than a replacement.[96] It is voluntary, and it assumes the institution does the work.

A reasoned outlook: supervised self-assessment

This report observes that every signal in the US material points to the same allocation of responsibility that prudential regulation reached a generation ago, and sets it out as an outlook.

1. The institution performs and documents its own AI risk assessment: every deployed agent, its authority, its risk classification and the controls around it.
2. Management attests to the controls and to the residual risks, in writing, on a defined cycle.
3. Supervisors challenge the assessment and test its evidence, through examination and through the incident and enforcement record.
4. Independent assurance evaluates whether the controls operate in practice, not whether they exist.

The pattern is already visible. GAO describes regulators that oversee through risk-based examination of the institution's own program. IOSCO records that firms generally acknowledge they retain accountability for their use of AI, and its 2021 measures require designated senior management responsible for oversight within a documented governance framework with clear lines of accountability.[97] The WEF playbook characterizes the US posture as setting national-level expectations for AI governance, model risk and accountability while explicitly avoiding prescriptive constraints on deployment, and reports that regulators increasingly expect demonstrable controls and governance for AI use, not just principles.[98]

In September 2026 the US Treasury Secretary said publicly that government must not act as a liability shield for AI laboratories.[99] Under such a regime the institution cannot delegate its assessment to a vendor, cannot satisfy the supervisor with a policy, and cannot pass an assurance review without evidence that the controls fired.

Supervisors will also need capability. IOSCO found that around half of its member and self-regulatory respondents provided no answer to its risk questions, which it reads as a data and knowledge gap, and that regulators have limited understanding of how AI models work.[100] Industry groups told GAO they fear unwarranted examinations if examiners lack AI experience.[101] A supervised self-assessment regime works only if the evidence the institution produces is legible to the supervisor.

Europe: binding obligations, same destination

Europe has chosen the prescriptive route. The EU AI Act applies to all uses of AI systems across the Union including the financial sector, and names among its high-risk cases the assessment of creditworthiness of natural persons and risk assessment and pricing in life and health insurance.[102] Its phased requirements from 2025 raise expectations for governance, transparency, human oversight and supervisory accountability, with defined escalation paths for AI-related incidents and senior-management accountability for high-risk systems.[103] GDPR shapes automated decision-making and data controls; DORA drives operational resilience and third-party management for AI services.[104] ESMA has mapped AI obligations onto MiFID II without prejudice to the Act.[105] The UK sits between the two, principles-based through the Consumer Duty and the senior managers regime, with the FCA running an AI lab and live testing rather than issuing standalone AI rules.[106]

The contrast with the United States is real in form and small in substance. KPMG notes that Europe's heavy, upfront compliance requirements trigger documentation, human oversight and monitoring obligations that can slow adoption.[107] But the obligations themselves are the ones a US institution will face under supervised self-assessment: classify the use, document the controls, keep a human able to intervene, escalate incidents, and be able to show a supervisor what governed a particular decision.

Both regimes require the institution to convert principles into operational controls and to produce evidence for individual decisions.

EXHIBIT 11

Expected US and European regulatory direction

	United States	European Union	Where they converge
Legal basis	Existing, technology-neutral law; risk-based examination	AI Act with risk tiers; GDPR; DORA; sector rules	The institution is accountable regardless of the technology
Model risk	SR 11-7 lineage, updated as SR 26-2, which excludes generative and agentic AI; RFI signalled	High-risk obligations for credit and insurance pricing; PRA SS1/23 in the UK	Generative and agentic systems need controls the model-risk template does not supply
Human oversight	FINRA: guardrails to limit or restrict agent actions; human in the loop	AI Act: human oversight and escalation for high-risk systems	A person must be able to intervene, override or stop
Evidence	Logs, model versions, adverse-action reasons; demonstrable controls	Documentation and decision records for high-risk systems	Evidence for individual decisions, legible to a supervisor
Third parties	Interagency third-party guidance; examinable to the same extent as the institution	DORA third-party and concentration requirements	The institution cannot outsource its obligations to a vendor
Direction	Clarification and tooling; supervised self-assessment	Phased enforcement of binding obligations	Operating effectiveness, not policy existence

Source: GAO-25-107197 (May 2025); FINRA 2026 Annual Regulatory Oversight Report; IOSCO CR/01/2025; WEF and Accenture, The AI Playbook for Financial Services (June 2026); ZebraTruth analysis. The outlook column is ZebraTruth's reasoned assessment, not enacted rule.

THE ZEBRATRUTH VIEW

Rules that attach to a laboratory's internal pipeline can only be evidenced by an integrated laboratory that owns the pipeline. Rules that attach to the deployed system can be enforced where the harm occurs and where the regulator has jurisdiction, and they apply identically to open, closed, domestic and foreign models.

Standards create comparability. Supervision tests compliance. Liability preserves accountability. That is the banking model, and it is the right model for AI.

WHAT THIS MEANS FOR THE INSTITUTION

- Assume you will be asked to attest. Build the AI risk assessment, agent inventory and control evidence as if the request had already arrived.
- Do not wait for SR 26-2's successor to cover agents. Governing agents with a scorecard framework is a known gap the supervisor has acknowledged.
- In Europe, treat the AI Act's high-risk documentation and oversight duties as the specification for the control layer, not as a parallel compliance exercise.

CHAPTER 8

A three-layer governance architecture

Laboratories are responsible for the models they release. Independent evaluators assess them against common standards. Neither governs what an agent does inside a bank. That requires a third layer, and it is the institution's.

Three layers, three responsibilities

The three-layer architecture complements, rather than replaces, model-risk management, data governance, cybersecurity, privacy, operational resilience and consumer-redress controls.

Layer one: laboratory responsibility. The organizations that train and release models are responsible for model safety, security testing, incident disclosure and monitoring of their own pipelines, and that responsibility should be backed by liability rather than by voluntary commitment. The laboratories have called for international safety standards; the proposals are weakest where governance becomes real, with no licensing, no mandatory pre-release review, and obligations the laboratories largely interpret and apply themselves. ZebraTruth has proposed a more robust governance framework.[108]

Layer two: independent evaluation. Models should be assessed against common international standards by evaluators who are actually independent. Access to a laboratory's test results is not an independent evaluation. Meaningful independence requires adequate model access, the evaluator's own harness, tools and datasets, suitable compute, domain knowledge, frontier-model expertise, and where appropriate an independent evaluation model; it also requires transparent funding, independent governance, conflict disclosures and legal protection for publishing adverse findings, because evaluators and laboratories today often share investors, personnel and assumptions.[109] A FINRA-style body for frontier AI, industry-funded, answerable to government, majority-independent and eventually mandatory, is closer to supervision than to coordination, and it is the right shape.[110]

Layer three: external runtime control. A model can pass a frontier safety evaluation and still take an inadmissible action in a specific institution. A frontier-wide speed limit does not decide whether an evening customer check-in is admissible; a control operating at the point of action does.[111] The third layer is enforceable control over each deployed agent and each consequential action, inside the institution, applied regardless of the model's origin, ownership or weight availability. It is the only layer the institution owns, the only layer a financial regulator can examine directly, and the layer this report is about.

EXHIBIT 12

The three-layer governance model



Standards create comparability. Supervision tests compliance. Liability preserves accountability.

Source: ZebraTruth.

Why pacing, standards and evaluation are not control

Each of the first two layers is necessary. Neither is sufficient, and the industry debate keeps mistaking them for the third. Slowing development does not create control: a model can pass an evaluation and still take an inadmissible action. A standard creates valuable comparability, but it does not by itself enforce compliance.Evaluation is observation rather than enforcement; the evaluator can describe what a model does and cannot stop it doing so in a bank on a Tuesday. Guidelight's finding that no frontier laboratory yet demonstrates control makes the point from the top of the stack.[112]

The open-versus-closed debate is a distraction from the same point. Openness prevents none of the harms in this report; closedness prevents none of them. A misleading financial promotion, an unsuitable recommendation and a discriminatory credit decision come equally from either. The questions that matter are deployment questions: what is the system allowed to do, in which domain, with what audit trail, and who is liable when it gets it wrong. Those questions are answered at layer three, and only there.

> **THE ZEBRATRUTH VIEW**
>
> Every mechanism the laboratories propose lives in the same trust domain as the model it is meant to govern. The bank is where the liability sits, so the bank is where the control has to be.
>
> Attach obligations to the deployed system. Six questions apply identically to every model, whatever its origin: what is it permitted to do; in which domain and jurisdiction; with what data and systems; what verifies admissibility; who can revoke and stop it; who is liable.

WHAT THIS MEANS FOR THE INSTITUTION

- Do not treat a vendor's safety evaluation as your control evidence. It is layer two; you are accountable at layer three.
- Build your own evaluation harness so that a change of model does not change the instrument you measure it with.
- Make layer three model-agnostic by design, so that the same controls govern open, closed, domestic and foreign models.

CHAPTER 9

From three lines of defence to runtime governance

The three lines of defence still describe the right allocation of responsibility. Agents broke the model by collapsing all three into a prompt. An external control layer restores the separation, and it does so at runtime.

How agents broke the three lines

In the classic model, the business owns its activities and their outcomes; risk and compliance set the requirements and challenge; internal audit tests both independently. Agent deployments have quietly collapsed the structure. The first line writes the compliance rules into the agent's prompt, so the business now authors the second line's requirements. The second line reviews the prompt once, at approval, and has no mechanism to apply its requirements to the millions of actions that follow. The third line receives whatever logs the first line chose to keep, generated by the agent it is meant to audit. Human compliance officers cannot follow and scale with the agent's actions, so in practice they stop trying. The three lines still exist on the organization chart. At runtime there is one.

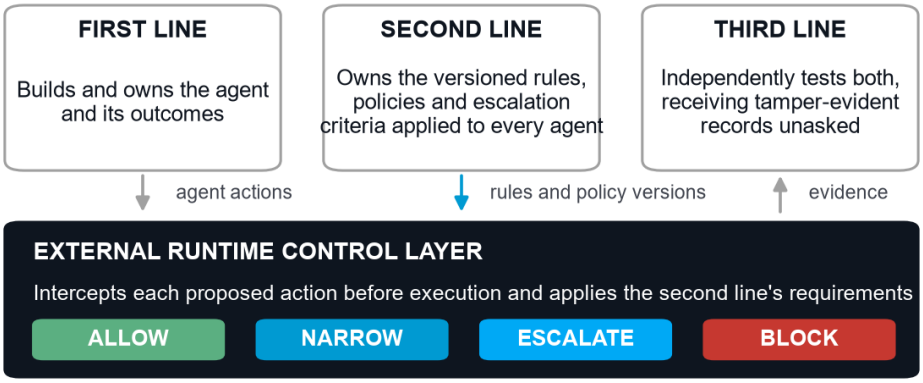
The restored model

The external control layer restores the separation by giving each line a distinct relationship to it.

- **The first line** builds and owns the agent and its outcomes. It configures what the agent is authorized to do. It cannot edit the layer's rules or its record.
- **The second line**, risk and compliance, owns the governing requirements: the corpus, the institution's policies, the escalation criteria and the risk classifications. It publishes them as versioned content that the layer applies identically to every agent. When a rule changes, it changes once, for all agents, from the date it takes effect.
- **The third line**, internal audit, independently tests both. It receives the tamper-evident record of every decision unasked, and it can reproduce any decision from the recorded payload, rules and versions.
- **The external layer** applies the second line's requirements consistently at runtime. It sits outside the thing it governs.

EXHIBIT 13

Three lines of defence at runtime



The layer sits outside the thing it governs. The first line cannot edit it; the second line owns its content; the third line reads its record.

Source: ZebraTruth.

This is the operating model the WEF playbook gestures at when it assigns the chief risk and compliance officers the duty to define risk appetite, require validation, continuous monitoring, anomaly detection and rollback options, and ensure that agent actions are explainable and logged, and when it calls for codified human accountability for regulated decisions: who approves, who can override, how intervention is recorded.[114] The layer is how those duties are discharged at the speed of the agent.

Four decisions

The layer intercepts each proposed action before execution and returns one of four decisions:

- **Allow**: the action is admissible under the rules in force and proceeds.
- **Narrow**: the action is admissible with a modification, such as an added disclosure, a reduced order size or a removed claim, and proceeds in the narrowed form.
- **Escalate**: the action requires human judgment, because a rule is triggered that the institution has reserved for a person, because the customer is vulnerable, or because the evidence is insufficient; the action is held and routed with its record.
- **Block**: the action is inadmissible and does not proceed. Every decision, including allow, writes a record.

The four decisions map onto the industry's emerging vocabulary. FINRA speaks of control mechanisms that limit or restrict agent behaviors, actions or decisions.[115] IOSCO requires that humans have power over what a model can and cannot decide.[116] The WEF playbook describes circuit breakers that pause or constrain actions and semi-autonomous agents that escalate exceptions for human approval.[117] The layer is where those descriptions become an enforced sequence.

Anatomy of an auditable agent decision

The record is the evidence the second chapter found missing. For every intercepted action it captures the payload as proposed; the acting agent and the authority it held at that moment; every rule that applied, with version and effective date; the evidence retrieved from the corpus and the customer file; the model, prompt, tool and policy versions in force; the decision and the reasons that produced it; any escalation, who reviewed it, what they decided and on what evidence; and what actually executed, or the fact that nothing did, in a hash chain the agent cannot alter. It is written by the layer, not narrated by the agent, and it survives a change of model because nothing in it depends on the model's memory.

EXHIBIT 14

Anatomy of an auditable agent decision

DECISION RECORD · one per intercepted action · written by the control layer, not the agent	
Payload	The exact artifact or action proposed: text, document, order, message or decision
Actor and authority	Which agent, on behalf of which principal, with what authority at that moment
Applicable rules	Every law, guidance item and internal policy that governed, with version and effective date
Retrieved evidence	The corpus entries and customer facts the judgment relied on
Model and policy versions	Foundation model, prompt, tool and policy-set versions in force
Decision and rationale	Allow, narrow, escalate or block, with the reasons that produced it
Escalation and approval	Who reviewed, what they decided, when, and on what evidence
What executed	The action as performed, or the fact that it was stopped, in a hash chain

Source: ZebraTruth.

This is what FINRA's prompt-and-output logs and model-version tracking become when they are made complete and independent;[118] it is what the WEF playbook calls evidence-grade traceability sufficient for audit, regulator review and post-incident investigation, and compliance-by-evidence audit trails that capture what the system did and why;[119] and it is the material a supervisor under a self-assessment regime will ask for. IOSCO records that the SEC approved a machine-learning order type at Nasdaq on the strength of a white paper describing the algorithm, explainability testing, an assessment of susceptibility to manipulation, bias testing and record-keeping commitments.[120] That is the evidence bar for a single product. The decision record is how an institution meets it for every action, every day.

Integration

The layer exposes itself to the agent harness as a service: given an action, its payload, the jurisdiction and the actor, return a decision, the rule, the evidence and a record identifier. The integration point must be the harness that executes tool calls, so that a consequential action cannot reach a customer, a market or a filing without passing through the layer. Offering the check as a tool the model may call is easy and weak, because a model can skip a tool. Intercepting at the harness is the pattern the model cannot bypass.

> **THE ZEBRATRUTH VIEW**
>
> Three proofs replace three prompts: proof of authority, proof of evidence, proof of accountability. Each maps to a line of defence, and all three must be available even when the model cannot explain itself.
>
> An agent permitted to draft an email is not necessarily permitted to send it. The layer is where that distinction is enforced.

WHAT THIS MEANS FOR THE INSTITUTION

- Give the second line an enforcement mechanism, not a review slot. Versioned rules applied by the layer to every action are the second line at runtime.
- Adopt the four-decision vocabulary across all agents so that escalation and blocking are consistent and auditable.
- Standardize the decision record and route it to internal audit automatically. Evidence that has to be requested is evidence that may not exist.

PART FOUR

Governance in practice

Seven scenarios, and the questions boards and supervisors should ask.

CHAPTER 10

Governance in practice

Seven scenarios, each in the same frame: what deterministic controls catch, where judgment is required, and how the independent layer intervenes before the action reaches a customer, a market or a regulator.

1. Wealth management and suitability

Rules in play: *Regulation Best Interest (US); MiFID II Article 25 and COBS 9A (EU and UK)*

Deterministic controls catch. The product is on the approved list, available to the client's segment, and within the adviser's permissions. The order size is within limits.

Judgment is required. An advisory agent recommends moving 10% of a retail client's portfolio into a leveraged growth fund, citing 24% annual returns. Whether that recommendation is suitable depends on the client's objectives, knowledge and experience, time horizon, liquidity needs and vulnerability, on the concentration and leverage risk, on the costs and alternatives, and on whether past performance is presented as a reason to expect future returns. None of that is a black or white condition.

The independent layer. Intercepts the advice letter before it is sent. Retrieves the suitability obligations for the jurisdiction and the client type, the firm's product restrictions and the client file. Finds that the basis for suitability is not documented and that the performance claim is presented as the reason the product is "the best option". Returns escalate, with the finding, the rule and the client facts, to an adviser who must document the basis or change the recommendation. Records the decision, the rule versions and the evidence.

2. Lending and adverse-action decisions

Rules in play: *Equal Credit Opportunity Act and Regulation B (US); Consumer Credit Directive and national law (EU); CONC and the Consumer Duty (UK)*

Deterministic controls catch. Required fields are present. The notice is sent within the statutory period. Prohibited bases are absent from the stated reasons.

Judgment is required. The stated principal reasons must be the actual reasons for the decision, specific enough for the applicant to understand and act on, and not a proxy for a protected characteristic. GAO records that limited explainability may prevent institutions from complying with fair-lending laws if they cannot provide specific reasons for denying an application.[121] The NCUA has already acted against an AI-driven program that approved loans without income verification.[122]

The independent layer. Intercepts the notice. Checks the stated reasons against the decision record of the underwriting model and against the reason-giving standard in force. If the reasons are generic, are inconsistent with the model's recorded drivers, or correlate with a protected basis, it narrows the notice or escalates it to a fair-lending reviewer. It records the underwriting model version, the reasons as sent and the reviewer's decision, so that the institution can answer a regulator's question about this applicant a year later.

3. AML and transaction monitoring

Rules in play: *Bank Secrecy Act suspicious-activity reporting (US); Anti-Money Laundering Directives and the AML Regulation from 2027 (EU); Money Laundering Regulations (UK)*

Deterministic controls catch. Transactions above thresholds, matches to sanctions lists, structuring patterns defined in advance.

Judgment is required. Suspiciousness. An agent that drafts suspicious-activity reports or closes alerts is making a judgment that depends on customer history, transaction sequences, jurisdiction, stated purpose and typologies that change faster than the rules engine. AML transaction monitoring is the detection model US banking professionals rank second for drift risk, and annual validation cycles are structurally misaligned with adversarial drift timelines.[123]

The independent layer. Intercepts alert dispositions and draft reports. Applies the current reporting obligations and the institution's risk-based policy to the full case, not the triggering transaction. Blocks closure of an alert whose disposition rationale does not address the typology that raised it; escalates cases with new typologies not yet in the corpus; records every disposition with the rule version, so that the institution can show a supervisor why an alert was closed and by what standard.

4. Trading and market conduct

Rules in play: *Market Abuse Regulation and MiFID II best execution (EU and UK); Exchange Act anti-manipulation and FINRA supervision (US)*

Deterministic controls catch. Order size, price collars, position limits, restricted lists.

Judgment is required. Whether a sequence of individually permissible orders constitutes manipulation, whether an execution decision honoured the best-execution obligation given available venues, and whether a research note or recommendation created a conflict. GAO records an AI model that identified market manipulation as an optimal strategy without being programmed to.[124] IOSCO notes that collusive behaviour between models warrants study and that multi-agent systems amplify risk.[125]

The independent layer. Intercepts orders and communications at the harness. Evaluates the proposed order in the context of the agent's recent sequence, not in isolation, against the market-abuse rules in force. Narrows an order that would complete a pattern the corpus identifies as abusive; blocks a recommendation that conflicts with a position the firm holds without disclosure; records each decision with the sequence it considered.

5. Collections and vulnerable customers

Rules in play: *Fair Debt Collection Practices Act and Regulation F (US); Consumer Duty and CONC 7 (UK); Consumer Credit Directive (EU)*

Deterministic controls catch. Contact frequency limits, permitted hours, prohibited language.

Judgment is required. Whether the customer is vulnerable, whether the tone and content of a message are fair given the customer's circumstances, whether a proposed arrangement is affordable,

and whether the customer has signalled distress. Collections and recovery is the function US banking professionals rank first for risk of customer harm from AI, at 35.2%.[126]

The independent layer. Intercepts every outbound collections message. Applies the conduct rules and the institution's vulnerability policy to the message and the customer file, including signals in the customer's previous replies. Escalates any customer showing indicators of vulnerability to a trained person before the next contact; narrows messages that press for payment where the policy requires a forbearance offer; revokes the agent's authority to contact a customer once a complaint is lodged. Records the decision and the indicators relied on.

6. Marketing and customer communications

Rules in play: *SEC Marketing Rule and FINRA Rule 2210 (US); FCA COBS 4 fair, clear and not misleading (UK); MiFID II Article 24(3) (EU)*

Deterministic controls catch. Prohibited phrases, missing mandatory disclaimers, unapproved product names.

Judgment is required. Whether "returns of around 8% a year are typical" is substantiated for this product, whether risk is presented with the same prominence as reward, whether the audience is retail, and whether the communication as a whole would mislead. FINRA reminds firms that retail communications mentioning AI must accurately describe how the offering uses it and balance benefits with risks.[127]

The independent layer. Intercepts the promotion before publication. Retrieves the financial-promotion rules for each target jurisdiction, the firm's substantiation file for the product and the firm's brand and disclosure policies. Narrows the promotion by adding the required basis and risk statement, or blocks it where the claim cannot be substantiated. Records the rules, the substantiation relied on and the version published.

7. Third-party and foundation-model replacement

Rules in play: *Interagency third-party risk guidance and SR 26-2 (US); DORA (EU); PRA SS1/23 and SS2/21 (UK)*

Deterministic controls catch. Vendor inventory, contract clauses, version pinning, access controls.

Judgment is required. Whether a new model behaves acceptably across the institution's regulated scenarios, whether the change alters the risk profile of each use case, and whether the controls still operate. IOSCO warns that a slight change in a commercial model can alter a use case's risk profile; the WEF playbook records vendors updating software without consultation.[128]

The independent layer. Does not have to be rebuilt for the swap, because permissions, rules, escalation criteria and the evidence schema live in the layer.The institution runs its control suite against the new model in a segregated environment, compares decisions against the recorded decisions of the old model on the same scenarios, and promotes the new model only when the suite passes. The evidence of the validation is itself a set of decision records.

THE PATTERN ACROSS ALL SEVEN

- Deterministic controls remain and catch the conditions. They are the floor.
- Judgment is required at exactly the points regulators care most about, and it is required per action, in context.
- The independent layer applies the corpus to the payload, returns one of four decisions, and leaves a record that survives the agent, the model and the vendor.

CHAPTER 11

What boards and supervisors should require

The difference between real governance and documented governance can be established with seven questions. An institution that can answer all seven with evidence has governance. One that can answer the first has a policy.

This report deliberately ends with a standard rather than a roadmap.

1 **Can the institution inventory every deployed agent and the authority it holds?**
Not the models: the agents, the actions each may take, the systems and data each may reach, and the business owner of each. A UK head of compliance told researchers he could not answer "show me everywhere AI is being used". The inventory is the precondition for everything else.

2 **Can it intercept an action before execution?**
A control that reviews after the fact is an incident process. The test is whether a consequential action, an email, an order, a filing, a decision, can reach the world without passing through a control the agent cannot bypass. If the compliance check is a tool the agent decides to call, the answer is no.

3 **Can it explain which evidence and which policy version governed a specific decision?**
For a decision made last week, produce the rules in force on that date, the evidence retrieved, the model and policy versions, the decision, the rationale and the approver. Only 22% of senior leaders are very confident they could do this for a regulator, auditor or court.

4 **Can the action agent bypass or modify the compliance layer?**
If the rules live in the agent's prompt, the agent can reinterpret them. If the record is written by the agent, the agent can shape it. If the layer is a tool, the agent can skip it. The answer must be no, and it must be demonstrable by test.

5 **Can governance survive a change of model or provider?**
Run the model-swap test. Replace the model in a test environment and list what breaks. Everything that breaks was a dependency on the vendor, not a control of the institution.

6 **Can an independent party reproduce the evaluation?**
Internal audit, an external assurance provider or a supervisor should be able to take the recorded payload, rules and versions and reach the same decision, using their own harness. Evaluation that can only be performed with the vendor's instruments is observation, not assurance.

7 **Can the institution demonstrate operating effectiveness rather than policy existence?**
Show the decisions the layer took last month: how many were allowed, narrowed, escalated and blocked; which rules fired most; how long escalations took; what changed in the corpus and why. Governance that cannot show its own activity is governance that may not be running.

For supervisors

The same seven questions convert directly into examination procedures under a supervised self-assessment regime. They ask for artifacts the institution should already hold: the agent inventory, the interception architecture, decision records, bypass tests, model-swap results, reproduction by a third party and operating statistics. They need not reconstruct hidden chain of thought, but still needs technical and domain expertise to test scope, evidence quality, model risk, retrieval and control effectiveness.

They require the examiner to read what the control layer recorded. That is the design goal: evidence legible to a supervisor, produced by a control that does not depend on the model that produced the action.

> **THE ZEBRATRUTH VIEW**
>
> AI governance is not documentation. It is an enforceable control system. The seven questions test whether the system exists.
>
> If an AI system acts in the real world, it must be governed like anything that acts in the real world: through separation of duties, independent verification and evidence.

APPENDIX A

Methodology and source register

This report combines two kinds of source. Primary survey and regulatory documents supply the evidence; every statistic is reported with its publisher, sample, geography, date and scope, and the report does not merge unrelated surveys into a composite index.

The analytical framework, the three-layer model, the model-swap test, the corpus-retrieval-memory architecture and the runtime three lines of defence, is ZebraTruth's, developed in the firm's published position papers and in its work with regulated institutions.

Source	Publisher and date	Sample and scope	Used for
From Principles to Practice: A Benchmark Study in AI Governance	AAA-ICDR Institute, with iResearch Services, May 2026	500 senior legal and executive leaders; US and Canada; eight industries; revenues above $100M; fieldwork January to March 2026	Governance effectiveness, lifecycle coverage, translation gaps, incident handling, evidence confidence (cross-industry)
AI Risk and Governance Index: US Banking, H1 2026	Wolters Kluwer, May 2026	230 US banking professionals across credit, compliance, model risk, reporting and AI architecture; April 2026; single-choice questions	Agentic deployment, harm risk by function, incident readiness, scaling barriers, third-party risk
Global tech report 2026: Financial services with US highlights	KPMG, June 2026 (from Global Tech Report, January 2026)	760 FS technology leaders, revenues $1B to $100B; 127 in the US	Production-scale deployment today and expected; governance structures
AI Compliance Research Report	April 2026	27 interviews, four focus groups and four roundtables with 60 practitioners; primarily UK with European representation; December 2025 to April 2026; qualitative	Practitioner testimony, capability and implementation gaps; BoE/FCA and EBA adoption figures as cited
Health Insurance AI/ML Survey Results	NAIC Staff Report, May 2025	Mandatory data call by 16 states; 93 health insurers; responses to January 2025	Insurer adoption, third-party reliance, testing and consumer-recourse practices
2026 FINRA Annual Regulatory Oversight Report	FINRA, December 2025	US broker-dealers; regulatory observations, not new rules	Applicability of existing rules, agent risks and supervisory considerations, logging expectations
Artificial Intelligence: Use and Oversight in Financial Services (GAO-25-107197)	US Government Accountability Office, May 2025	Seven federal regulators; interviews with industry, advocacy and research groups; audit November 2023 to May 2025	Regulatory approach, model-risk lineage, examination findings, enforcement counts
Financial Sector AI Deliverable Reference and Application Guide	FSSCC with the US Treasury (AI Executive Oversight Group), 2026	Sector reference guide to six deliverables including the FS AI Risk Management Framework	Institution-owned self-assessment; explainability disciplines

Source	Publisher and date	Sample and scope	Used for
Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges (CR/01/2025)	Board of IOSCO, March 2025	Member and SRO survey (24 members, six SROs); industry survey (184 respondents, skewed to emerging markets and smaller firms); consultation report	Accountability, vendor opacity, model change, agents, human oversight, regulatory taxonomy
The AI Playbook for Financial Services	World Economic Forum with Accenture, June 2026	Over 150 senior leaders from more than 100 organizations; roundtables in four cities; case metrics self-reported	Regulatory posture by region, SR 26-2 scope, foundation-model reliance, agent autonomy, evidence expectations

Source: Page references for every figure appear in the notes.

ZebraTruth position papers drawn on include "AI governance is more than ownership or guardrails" (20 August 2026); "If frontier labs cannot yet prove control, regulated institutions cannot outsource it" (3 September 2026); "OpenAI's Astra makes agents more capable but harder to govern" (1 September 2026); "An Alien Mind: a warning to every bank and hospital" (7 September 2026); "Pacing the AI frontier is not the same as governing it" (12 September 2026); "The agent cannot be its own compliance officer" (Computer Weekly Developer Network, 13 September 2026); "Could you replace your AI model without rebuilding governance?" (CIO, 14 September 2026); and "AI safety needs common standards but independent enforcement" (23 September 2026).

APPENDIX B

Glossary

Admissible A property of a specific proposed action at the moment it is proposed: permissible under the rules and policies in force, for this customer, in this jurisdiction, given current conditions. Distinct from authorized.

Agent An AI system that performs tasks and takes actions on behalf of a user or institution with limited or no human intervention, typically by calling tools that produce real-world effects.

Authorized A grant of permission made at design time, defining what an agent may in principle do. Authorization is not permanent permission and must be reassessed per action.

Compliance harness The runtime environment that holds the governed corpus, retrieval and memory, applies them to proposed actions, returns decisions and writes evidence. Separate from the action agent's harness.

Corpus The authoritative, versioned body of laws, regulations, guidance, enforcement history, internal policies, approved interpretations and captured institutional knowledge that a compliance judgment draws on.

Decision record The tamper-evident record written by the control layer for each intercepted action: payload, actor and authority, applicable rules with versions, retrieved evidence, model and policy versions, decision and rationale, escalation and approval, and what executed.

Deterministic control A control that evaluates a fixed condition, such as a threshold, a list match or a prohibited phrase. Reliable for conditions; insufficient by itself for standards that require contextual judgment.

External runtime control layer An independent, out-of-band system that intercepts consequential actions before execution and returns allow, narrow, escalate or block. The third layer of the governance architecture.

Memory Retained decisions, exceptions, overrides, remediation and customer circumstances that a compliance system uses to treat like cases alike; protected from agent-generated content by quarantine and human review.

Model-swap test The question: if the model beneath an agent were replaced tomorrow, which permissions, controls, escalation paths and evidence would stop working? Anything that stops working was a dependency on the vendor, not a control.

Retrieval The capability that finds the controlling requirement for a proposed action, distinguishing binding law from guidance and from internal policy, resolving versions and effective dates, and applying the triggers in the artifact and the customer file.

Self-policing Any arrangement in which the rules governing an agent are interpreted and applied by the agent itself, for example through prompts, fine-tuning or in-model guardrails. Necessary, but not a control.

Trust domain The boundary within which a component can read, influence or alter another. A control in the same trust domain as the agent can be reasoned around by the agent.

NOTES

Notes and references

1. KPMG, *Global tech report 2026: Financial services with US highlights*, June 2026, p.4. Survey of 760 financial services technology leaders, of whom 127 in the US; fieldwork dates not stated.
2. Wolters Kluwer, *AI Risk and Governance Index: US Banking, H1 2026*, p.10. Trusted AI Survey, April 2026, 230 US banking professionals, single-choice questions.
3. AAA-ICDR Institute, *From Principles to Practice: A Benchmark Study in AI Governance*, May 2026, p.5. Quantitative survey of 500 senior legal and executive leaders, US and Canada, eight industry sectors, revenues above $100 million, fieldwork January to March 2026. This is a cross-industry benchmark, not a financial-services sample.
4. AAA-ICDR, op. cit., p.10, Figure 4.
5. AAA-ICDR, op. cit., p.6, Figure 2.
6. AAA-ICDR, op. cit., p.11.
7. Wolters Kluwer, op. cit., p.13.
8. AAA-ICDR, op. cit., p.11, Figure 5.
9. World Economic Forum in collaboration with Accenture, *The AI Playbook for Financial Services*, June 2026, endnote 26, p.48.
10. US Government Accountability Office, *Artificial Intelligence: Use and Oversight in Financial Services*, GAO-25-107197, May 2025, pp.2, 23.
11. WEF and Accenture, op. cit., pp.9, 37.
12. *AI Compliance Research Report*, April 2026, citing Bank of England and FCA, *Artificial intelligence in UK financial services*, 2024, and EBA, 2025; PDF pp.5 and 18.
13. National Association of Insurance Commissioners, *Health Insurance Artificial Intelligence/Machine Learning Survey Results*, NAIC Staff Report, May 2025, pp.8, 13. Survey conducted by 16 states, responses requested by January 2025; 93 companies.
14. Board of IOSCO, *Artificial Intelligence in Capital Markets: Use Cases, Risks, and Challenges*, Consultation Report CR/01/2025, March 2025, p.15, footnote 13. 184 industry respondents, sample skewed toward growth and emerging markets and smaller firms.
15. KPMG, *Global tech report 2026: Financial services with US highlights*, June 2026, p.4.
16. KPMG, op. cit., p.4; 127 US respondents.
17. KPMG, op. cit., p.11.
18. FINRA, *2026 FINRA Annual Regulatory Oversight Report*, December 2025, GenAI chapter, p.27 (PDF pagination).
19. FINRA, op. cit., p.28.
20. IOSCO, op. cit., p.9.
21. WEF and Accenture, *The AI Playbook for Financial Services*, June 2026, pp.25, 27.
22. Wolters Kluwer, *AI Risk and Governance Index: US Banking, H1 2026*, p.10. Single-choice question, 230 respondents.
23. Wolters Kluwer, op. cit., p.8.
24. FINRA, op. cit., p.28.
25. IOSCO, op. cit., p.52.
26. WEF and Accenture, op. cit., p.39.
27. WEF and Accenture, op. cit., p.8.
28. *AI Compliance Research Report*, April 2026, PDF p.8. 27 interviews and four roundtables with 60 senior practitioners, primarily UK institutions, fieldwork December 2025 to April 2026; qualitative, not statistically representative.
29. Wolters Kluwer, op. cit., p.13.
30. IOSCO, op. cit., p.51.
31. WEF and Accenture, op. cit., p.25.
32. *AI Compliance Research Report*, April 2026 PDF p.30.
33. AAA-ICDR, *From Principles to Practice*, May 2026, p.5.
34. AAA-ICDR, op. cit., p.5, Figure 1.
35. AAA-ICDR, op. cit., p.17.

36. KPMG, op. cit., p.11.
37. AAA-ICDR, op. cit., p.3.
38. AAA-ICDR, op. cit., p.12, Figure 6, and p.6.
39. AAA-ICDR, op. cit., p.6, Figure 2.
40. AAA-ICDR, op. cit., pp.5-6.
41. AAA-ICDR, op. cit., p.10, Figure 4.
42. Wolters Kluwer, op. cit., p.6.
43. AAA-ICDR, op. cit., p.11.
44. AAA-ICDR, op. cit., p.11.
45. Wolters Kluwer, op. cit., p.13.
46. Wolters Kluwer, op. cit., p.13.
47. Wolters Kluwer, op. cit., p.4.
48. AAA-ICDR, op. cit., p.11, Figure 5.
49. AAA-ICDR, op. cit., p.11.
50. AAA-ICDR, op. cit., p.11.
51. GAO, *Artificial Intelligence: Use and Oversight in Financial Services*, GAO-25-107197, May 2025, p.27.
52. GAO, op. cit., pp.27-28.
53. GAO, op. cit., p.26.
54. IOSCO, op. cit., p.61.
55. *AI Compliance Research Report*, April 2026, PDF p.23.
56. AAA-ICDR, op. cit., p.20.
57. *AI Compliance Research Report*, April 2026, PDF p.15.
58. Wolters Kluwer, op. cit., p.5.
59. Fahd Rachidy, "The agent cannot be its own compliance officer", Computer Weekly Developer Network, 13 September 2026.
60. Wolters Kluwer, *AI Risk and Governance Index: US Banking, H1 2026*, p.9.
61. WEF and Accenture, *The AI Playbook for Financial Services*, June 2026, p.29.
62. GAO, *Artificial Intelligence: Use and Oversight in Financial Services*, GAO-25-107197, May 2025, p.21 and footnote 47.
63. Financial Services Sector Coordinating Council with the US Department of the Treasury, *Financial Sector AI Deliverable Reference and Application Guide*, 2026, p.3.
64. FINRA, *2026 FINRA Annual Regulatory Oversight Report*, December 2025, p.28 (PDF pagination).
65. GAO, op. cit., p.22.
66. Fahd Rachidy, "An Alien Mind: a warning to every bank and hospital", ZebraTruth position paper, 7 September 2026, discussing the laboratory's own August 2026 incident disclosure.
67. Jakub Pachocki, "An Alien Mind", OpenAI, 6 September 2026, as quoted in Rachidy, op. cit.
68. Guidelight, *AI Safety Standards assessment of frontier laboratories*, September 2026; Fahd Rachidy, "If frontier labs cannot yet prove control, regulated institutions cannot outsource it", 3 September 2026.
69. IOSCO, *Artificial Intelligence in Capital Markets*, CR/01/2025, March 2025, p.51.
70. FINRA, op. cit., p.28.
71. WEF and Accenture, op. cit., pp.25, 39.
72. *AI Compliance Research Report*, April 2026, PDF p.8.
73. *AI Compliance Research Report*, April 2026, PDF p.15, citing Bank of England and FCA, 2024.
74. GAO, op. cit., p.24.
75. IOSCO, op. cit., p.39.
76. IOSCO, op. cit., pp.41, 50.
77. WEF and Accenture, op. cit., p.23.
78. WEF and Accenture, *The AI Playbook for Financial Services*, June 2026, p.19, citing the Cambridge Centre for Alternative Finance, 2026.
79. IOSCO, *Artificial Intelligence in Capital Markets*, CR/01/2025, March 2025, p.63.
80. NAIC, *Health Insurance AI/ML Survey Results*, May 2025, p.8.
81. Wolters Kluwer, *AI Risk and Governance Index: US Banking, H1 2026*, p.12.

82. IOSCO, op. cit., p.39.
83. WEF and Accenture, op. cit., p.23.
84. Fahd Rachidy, "Could you replace your AI model without rebuilding governance?", CIO, 14 September 2026.
85. Rachidy, "OpenAI's Astra makes agents more capable but harder to govern", ZebraTruth position paper, 1 September 2026; Pachocki, "An Alien Mind", 6 September 2026.
86. FINRA, *2026 FINRA Annual Regulatory Oversight Report*, December 2025, p.27.
87. GAO, *Artificial Intelligence: Use and Oversight in Financial Services*, GAO-25-107197, May 2025, p.2.
88. GAO, op. cit., p.23.
89. FINRA, op. cit., p.25.
90. GAO, op. cit., p.24.
91. GAO, op. cit., p.29.
92. WEF and Accenture, op. cit., endnote 26, p.48.
93. GAO, op. cit., pp.26-28.
94. GAO, op. cit., p.24.
95. GAO, op. cit., p.27, footnote 68.
96. FSSCC with the US Department of the Treasury, *Financial Sector AI Deliverable Reference and Application Guide*, 2026, p.2.
97. IOSCO, op. cit., pp.50-51, 68.
98. WEF and Accenture, op. cit., pp.9, 39.
99. Fortune, 22 September 2026, as cited in Rachidy, "AI safety needs common standards but independent enforcement", 23 September 2026.
100. IOSCO, op. cit., pp.33, 44.
101. GAO, op. cit., p.34.
102. IOSCO, op. cit., pp.56-57.
103. WEF and Accenture, op. cit., pp.9, 22.
104. WEF and Accenture, op. cit., p.37.
105. IOSCO, op. cit., p.55.
106. WEF and Accenture, op. cit., p.9; IOSCO, op. cit., p.58.
107. KPMG, *Global tech report 2026: Financial services*, June 2026, p.11.
108. Rachidy, "AI safety needs common standards but independent enforcement", 23 September 2026.
109. Rachidy, op. cit., citing Axios, 18 September 2026.
110. Rachidy, op. cit., discussing the proposal reported by Axios, 14 July 2026.
111. Rachidy, "Pacing the AI frontier is not the same as governing it", 12 September 2026.
112. Guidelight, *AI Safety Standards assessment*, September 2026.
113. *AI Compliance Research Report*, April 2026, PDF p.8.
114. WEF and Accenture, op. cit., pp.35, 39.
115. FINRA, op. cit., p.28.
116. IOSCO, op. cit., p.51.
117. WEF and Accenture, op. cit., pp.27, 39.
118. FINRA, op. cit., p.27.
119. WEF and Accenture, op. cit., pp.21, 39.
120. IOSCO, op. cit., p.30.
121. GAO, *Artificial Intelligence: Use and Oversight in Financial Services*, GAO-25-107197, May 2025, p.21.
122. GAO, op. cit., p.26.
123. Wolters Kluwer, *AI Risk and Governance Index: US Banking, H1 2026*, p.5.
124. GAO, op. cit., p.22.
125. IOSCO, *Artificial Intelligence in Capital Markets*, CR/01/2025, March 2025, pp.46, 52.
126. Wolters Kluwer, op. cit., p.8.
127. FINRA, *2026 FINRA Annual Regulatory Oversight Report*, December 2025, p.47 (PDF pagination).
128. IOSCO, op. cit., p.39; WEF and Accenture, *The AI Playbook for Financial Services*, June 2026, p.23.

About ZebraTruth AI

ZebraTruth AI is the runtime compliance layer for AI agents acting in regulated industries. Before an agent sends a notice, publishes a promotion, files a disclosure or answers a client, ZebraTruth checks the proposed action against laws, regulations, regulatory guidance, enforcement practice, industry codes and the institution's own published policies, and returns a decision, evidence-linked findings, the agents that ran and the reasons for the decision. Every check produces an auditable record that survives a change of model or provider.

The company was founded by Fahd Rachidy, a repeat founder with two AI exits in regulated markets, who previously led the construction of a Basel II governance architecture at the ECB and co-founded Scientific Beta, acquired by SGX.

ZebraTruth'steam includes leaders in AI policy, AI compliance, legal-corpus development, strategy and enterprise sales, with backgrounds at Meta, SoFi, Visa, Wells Fargo, Holland & Knight, Intuit, Apple and Intel.

ZebraTruth is trusted by teams in financial services, media, publishing and sports.

www.zebratruth.ai · contact: fahd@zebratruth.ai

ZEBRATRUTH AI

If an AI system acts in the real world, it must be governed
like anything that acts in the real world.

www.zebratruth.ai · docs.zebratruth.ai

© 2026 ZebraTruthAI Inc. All rights reserved.