The Center for
AI Oversight

CENTER PAPER NO. 5 · JUNE 2026

Agentic AI and the Oversight Lens

Governing systems that act. What changes when AI stops advising and starts doing, and why the oversight layer is the only framework that currently reaches it.

By **Brian J. Allen**, Executive Director

> *Each section moves in three tiers: a **board line**, a **why it matters** note, and a shaded **detail** block carrying the authority.*

EXECUTIVE SUMMARY

When AI acts, the program is the only live framework

KEY TAKEAWAYS

1. Agentic AI takes actions rather than producing outputs, and that single difference stresses every element of the oversight program: decision rights, incident definitions, and escalation tempo.
2. The banking agencies just rewrote the most mature model-governance framework in American supervision and deliberately scoped generative and agentic AI out of it. The oversight layer is not one framework among several for agents. It is the only live one.
3. Three things change when software acts: authority (the institutional act occurs without a contemporaneous human decision), attribution (the accountable chain multiplies), and tempo (the agent acts faster than any meeting).
4. The program's questions do not change, only the answers. The inventory must capture authority, not just software; a named human must own every agent; escalation triggers must be set before deployment; and the action log becomes the new board minute.
5. Closing the Velocity Gap for agents requires deciding the governance in advance. The institutions that do will deploy faster and be the only ones able to prove an informed decision was possible. The program is the governance.

This paper applies the Center's thesis to its hardest case: govern the system, not the technology. Agentic capability is exactly the technology that will not hold still, which is why oversight built around the technology fails and oversight built around the program endures.

This paper uses a deliberately operational definition of agentic AI and sources its claims about enterprise deployment to primary disclosures and named vendor documentation rather than press characterization.

PART ONE

What changes when AI acts

An agentic system, given an objective, takes sequences of actions with limited human review of the individual steps. Three things change when the institution's software stops recommending and starts doing: authority, attribution, and tempo.

Why it matters. Whether the underlying model is impressive is not the governance question. The governance question is what changes when the act occurs without a contemporaneous human decision, and each change maps to a program element that must be rebuilt for systems that act.

> AUTHORITY
> A recommendation borrows the authority of the human who accepts it; an agent exercises delegated authority directly. Delegation is the oldest subject in governance, and institutions have mature instruments for delegating to people, scope, limits, supervision, revocation, but almost none have written the equivalent for software. The agent's effective authority becomes whatever its technical permissions allow, and permissions were never designed to express governance intent.

AttributionWhen an agent's action goes wrong, attribution runs through a chain the institution may not have documented: who set the objective, who approved deployment, who defined the limits, who was monitoring, who could have stopped it. Agents do not remove the named-accountable-human requirement; they multiply the points at which the name must be recorded.

TempoAgents act at software speed, compound their own actions, and can spawn subordinate processes. Oversight designed around quarterly reporting is structurally mismatched to a system that can take a thousand actions between board meetings. The resolution is not faster meetings; it is moving the governance decisions forward in time.

PART TWO

The program, stress-tested

The oversight program's elements are stable across technologies, and agentic AI is the strongest test of that claim. In each case the question survives and the answer changes.

2.1 Inventory: systems that multiply

Inventory at the level of durable governance objects: the agent's identity, its objective and scope of authority, the systems it can touch, the human accountable, and its kill conditions.

Why it matters. Agents are configured rather than installed, call tools and other models, and can initiate subordinate processes. An inventory that cannot answer what an agent could do at maximum permission has captured the software and missed the authority, which is what the board is accountable for. Agentic capability is also arriving primarily through vendors, so the Buyer-side inventory discipline applies with the dial turned up.

2.2 Accountability: a named human for every agent

Single-point accountability per agent: one named owner who holds the delegation instrument, sees the action logs, owns the limits, and carries the upward duty, whatever domains the agent crosses.

Why it matters. McDonald's put officers under personal oversight duties within their domains, and a single agentic workflow crosses several. Shared accountability for an agent is unaccountability with extra names. The delegation instrument is the accountability question's new artifact, and boards should expect to see it the way they expect a committee charter.

2.3 Escalation: triggers before deployment

Triggers must be defined at deployment: action thresholds, anomaly conditions, out-of-scope attempts, and the stop conditions under which the agent halts pending human review.

Why it matters. A trigger invented after the incident is evidence for the plaintiff, and with agents the incident will be in progress when discovered. Incident definitions need rewriting in the same pass, because the agent's characteristic failure is misdirected competence, performing correctly toward the wrong objective, which a taxonomy built for model error will classify as success.

2.4 The record: action logs as the evidentiary core

The agent's action log, what it did, under what authority, within what limits, with what monitoring and interventions, is simultaneously the operational telemetry, the examination artifact, and the litigation exhibit.

Why it matters. The Informed Decision Standard asks the institution to prove informed decisions were possible. For agents the proof has a new center of gravity. The oversight obligations are retention, integrity, ownership, and reviewability of the logs; the EU AI Act's deployer log-retention duties and the state documentation defenses point the same way. An institution whose agents act without durable logs has built the Boeing record in machine time.

The agent's action log is the new board minute. It is the record that proves, or disproves, that the institution governed what acted in its name.

QUESTIONS FOR THE BOARD

- For every agent we deploy, is there one named owner, a written delegation instrument, and a durable action log?
- Were the stop conditions and escalation triggers defined before deployment, or do they not yet exist?

PART THREE

The open lane, and who must fill it

The scoping decisions of 2026 leave exactly one framework governing what an institution must oversee when software acts on its behalf: the institution's own oversight program.

Why it matters. Banking institutions cannot answer the agent question by pointing to model risk compliance; the agencies declined to let them. The frontier statutes bind the Builders of the most capable models, not the deploying institution's delegation of authority. The technical safety literature addresses the how layer, indispensable and insufficient. What remains is the oversight layer, and that is the assignment, not a gap to lament.

> **THE SCOPING RECORD**
>
> SR Letter 26-2 (Apr. 17, 2026) expressly placed generative and agentic AI outside the revised model risk framework. The Executive Order on advanced AI (June 2, 2026) addresses advanced systems at the national-security layer, not the institutional-oversight layer. Frontier statutes (Cal. Bus. & Prof. Code § 22757.10 et seq.; New York RAISE Act) bind developers; their published frameworks are Buyer diligence material, not governance of the deploying institution's agents.

The institutions that move fastest on agentic AI over the next two years will be the ones whose delegation instruments, escalation triggers, and log disciplines were decided in advance, because those are the institutions that can say yes to the next agent in days rather than quarters and defend the yes in any forum. The cyber-extension habit, stretching existing programs over agents, fails here for the same reason it fails for AI generally, and faster: the inherited program was built for systems that hold still and advise. Govern the system, not the technology, was written for exactly this moment.

CONCLUSION

The questions held; the program held

Agentic AI does not change the questions of oversight governance, and that stability is the most useful fact a board can hold as the technology accelerates.

What is governed: the inventory now includes authority, not just software. Who is accountable: a named human per agent, holding a written delegation. What escalates: triggers and stop conditions decided before deployment, on clocks the law already runs. What record exists: action logs kept with the discipline of board minutes, because that is what they now are. The institutions that answer those questions in advance will deploy agents faster than the institutions that do not, and will be the only ones able to prove, when the first agentic incident reaches a courtroom or an examiner, that an informed decision was possible. The program is the governance.

Sources and Further Reading

Authority: SR Letter 26-2, Revised Guidance on Model Risk Management (Federal Reserve, OCC, FDIC, Apr. 17, 2026) (scoping generative and agentic AI out); Executive Order on advanced AI innovation and security (June 2, 2026); Regulation (EU) 2024/1689, Arts. 14 and 26 (human oversight and deployer duties); Cal. Bus. & Prof. Code § 22757.10 et seq.; New York RAISE Act; SEC Release No. 33-11216, Form 8-K Item 1.05; In re McDonald's Corp. Stockholder Derivative Litig., 289 A.3d 343 (Del. Ch. 2023); In re The Boeing Co. Derivative Litig., C.A. No. 2019-0907-MTZ (Del. Ch. Sept. 7, 2021); NIST AI RMF.

The doctrinal framework appears in full in the Center's *Caremark AI Liability Roadmap*, and the Buyer-side discipline in *Builders and Buyers*. Current status of every obligation cited is maintained in *The AI Oversight Obligations Reference*.