

The Oversight Layer of AI Governance

Charter the oversight program. Everything runs from it.

Brian J. Allen, Esq., Executive Director, Center for AI Oversight. This paper states the Center's general position on the oversight layer. Sector applications, including the Center's comments to standards bodies, apply the same structure to specific instruments.

THE BOTTOM LINE

Three layers operate in any organization deploying AI: oversight governance, management, and technical controls. The lower two have owners and mature external instruments, the NIST and ISO standards and their sector equivalents. The oversight layer, which decides what is governed, at what granularity, and who is accountable, is usually missing, and nothing beneath it supplies it.

Seven elements charter that layer. Each is drawn from existing authority, none requires restructuring the layers below, and three of the seven carry most of the value for a resource-constrained institution.

A bank deploys a model that declines loan applications without a person reviewing each one. Ask who authorized it to act on its own, when that authority expires, and who can withdraw it. In most institutions the controls around the model are in good order and no one can answer. That is not a control failure. It is a missing layer.

Technical controls govern how the system is built, constrained, and watched. **Management** governs how the organization pursues its AI objectives and manages the risk of doing so. **Oversight governance** sits above both. It decides what is governed, at what granularity, and who is accountable when the decision is examined, and it never decides how.

The first two layers are well served. Engineering builds and operates the technical controls. The management layer is what ISO/IEC 42001:2023 is written for: a management system standard covering objectives, planning, resourcing, operation, and improvement, with the NIST AI Risk Management Framework (AI RMF 1.0) supplying the risk practice within it. The third layer is the one most institutions have not built, and its first decision is about those instruments themselves: whether to apply them, how far, and who decided.

Management names an activity, not a rank. Setting objectives, executing strategy, prioritizing, resourcing, and treating risk are management wherever they are performed, including at the top of the organization. A chief risk officer conducting a risk assessment is doing management. The same officer, named as accountable for the AI oversight program, holds an oversight position. What makes a decision oversight is not the seniority of the person making it. It is that the decision settles what is governed and who answers.

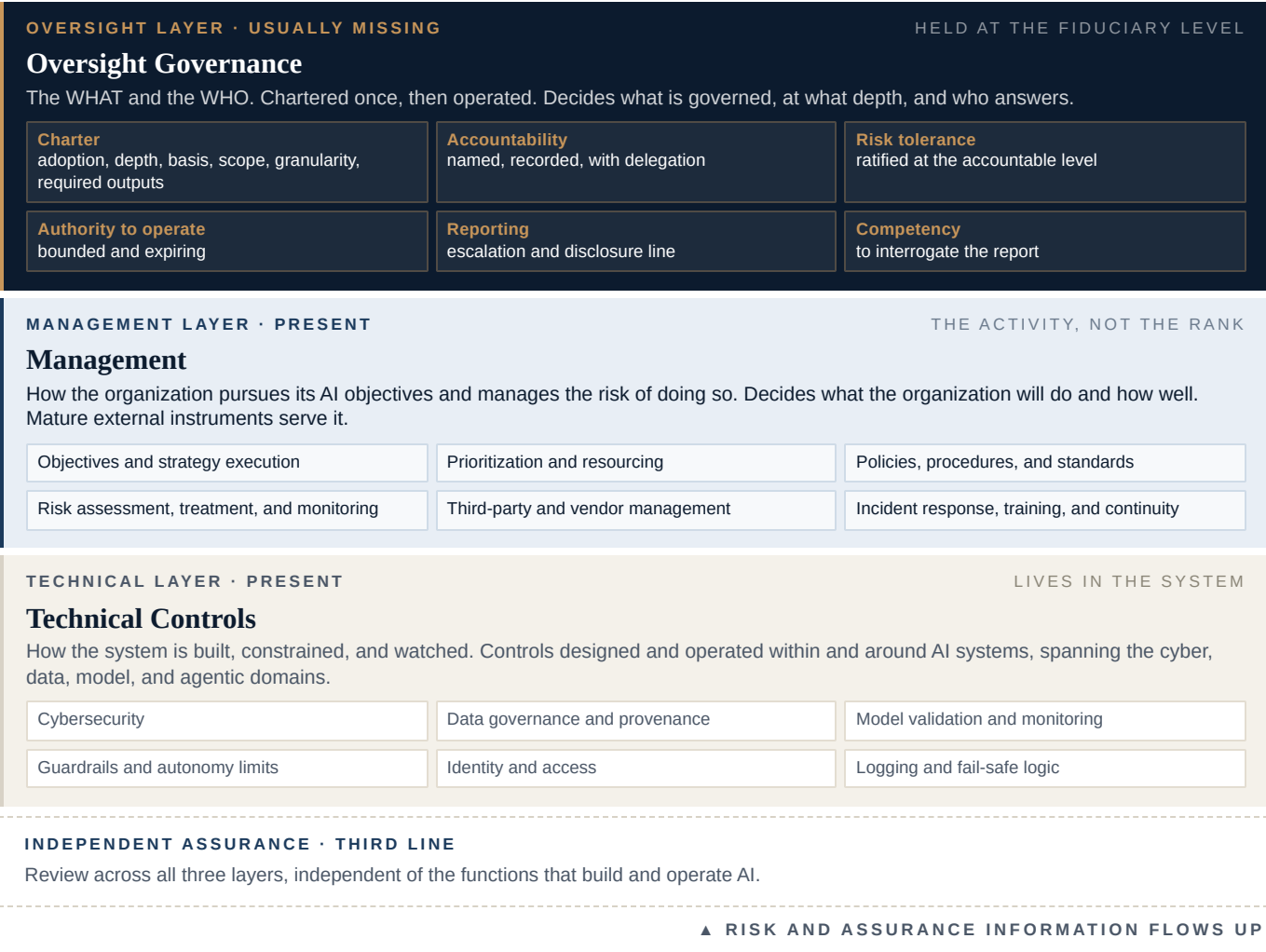


FIGURE 1 · THE LAYERS OF AI GOVERNANCE. Layer assignment follows subject matter, not audience altitude. Management names the activity, not the rank: when the chief executive sets an AI objective or a priority, that is management; when the same executive names who answers for a deployment, that is oversight. A control can be pitched at the C-suite and still describe only what the organization does. The charter is decided at the oversight layer and operated at the management layer beneath it.

Why existing practice does not supply the layer

GRC ASSIGNS ACTIVITY, NOT ANSWERABILITY

Governance, risk, and compliance practice produces control inventories, risk registers, policy sets, and testing programs, each with an owner. It rarely states which layer a responsibility belongs to, or who is answerable at the top of each.

The layers collapse into one undifferentiated function. The organization can then describe in detail how AI risk is managed without being able to say who is accountable for a given deployment.

THE INSTRUMENTS DEFER THE CALIBRATION

The management instruments correctly leave scope, granularity, and reporting depth to the organization: testing proportionate to consequence, review calibrated to criticality and risk tolerance.

Each is a calibration they are right to leave to the operator. None says who holds it, or on what record. A calibration nobody ratified is a working assumption, and an examiner cannot tell the two apart.

A committee is evidence of attention, not of accountability.

Six questions only the oversight layer answers

The duty of oversight, as the courts have framed it since *Caremark*, asks whether the accountable body made a good faith effort to put a reasonable system of monitoring and reporting in place. For AI, these are the six questions that system has to answer, and they are asked of the governing body and the executives accountable to it. Every control in the two layers beneath can be in place and none of them answers these, because each is a decision about accountability, not a control.

i THE PROGRAM Have we chartered an AI oversight program, and does the record show which standards we elected to apply, how far, and who decided?	ii THE ACCOUNTABILITY Which committee holds oversight of the program by charter, which officer answers to it by name, and what have we allowed to sit outside it, accepted by whom?	iii THE TOLERANCE Did we set the boundaries of acceptable AI risk in advance, on a stated basis, and ratify them, or were they set by default?
iv THE AUTHORITY What operates without a person in the loop, under what limits, until when, and who can withdraw it?	v THE LINE Do escalation and disclosure reach us on triggers and thresholds fixed in advance, by at least one path that does not depend on management's discretion, and is our position toward regulators set before an event rather than after?	vi THE RECORD When a regulator, court, or auditor asks, can we show the informed decision, the report we received, what we did with it, and that someone independent confirmed the program did what it said?

The oversight charter is what makes them answerable on a record. It is one page, amended onto a document the institution already keeps, and the seven elements set out later in this paper are how it is built. Question i is the charter itself (element 1). Question ii is named accountability at two levels, the committee and the officer (element 2). Question iii is ratified tolerance (element 3). Question iv is a bounded, expiring authority to operate (element 4). Question v is the reporting, escalation, and disclosure line (element 5). Question vi is the record the program produces, independent assurance over it, and the competency to act on what it reports (elements 6 and 7). Beneath all six, the program answers the same questions for each deployment: what it puts at risk, who is accountable for it, under what authority it runs, what would cause its withdrawal, and how the institution would learn that it should be.

The last time this layer went unbuilt, the gap took a decade to close

This is not the first control discipline to mature from the bottom up. The 2014 Cybersecurity Framework organized its core around five functions and contained no GOVERN function. The control layers matured first, the oversight layer went largely unbuilt, and the gap was closed a decade later from three directions at once: NIST added GOVERN as a sixth function in Cybersecurity Framework 2.0 in February 2024; the Securities and Exchange Commission adopted cybersecurity disclosure requirements in July 2023, including board oversight at Item 106(c)(1) of Regulation S-K; and in the courts the oversight duty of *In re Caremark Int'l Inc. Derivative Litig.*, 698 A.2d 959 (Del. Ch. 1996), was applied to mission critical risk in *Marchand v. Barnhill*, 212 A.3d 805 (Del. 2019), and to cybersecurity in *Firemen's Ret. Sys. of St. Louis v. Sorenson*, C.A. No. 2019-0965-LWW (Del. Ch. Oct. 5, 2021), where a documented and monitored program earned dismissal after a breach of up to 500 million guest records.

None of that displaced the control frameworks. It added a layer above them, after the fact, on terms the standards community did not set. What it produced at the receiving end is the part worth noting. Asked what their own oversight role is, operators, executives, and directors point to the Cybersecurity Framework. It is the artifact everyone reaches for, and it does not answer that question, because it was not written to. There has been nothing to point to, and that absence is the argument for populating this layer now rather than later.

Where the lines fall

Two tests place any control in its layer. Neither asks who performs it.

OVERSIGHT AGAINST MANAGEMENT

Management decides what the organization will do and how well. Oversight decides who answers for it, under what authority, within what tolerance, and on what record.

Setting an AI objective is management. Ratifying the tolerance the objective must stay inside is oversight. Deciding to deploy is management. Granting a bounded, expiring authority to operate is oversight.

MANAGEMENT AGAINST TECHNICAL CONTROLS

If the control changes what a person or a function must do, it is management. If it changes what the system does, it is a technical control.

The policy that requires a guardrail is management. The guardrail is a technical control. The procedure for invoking an override is management. The override mechanism is a technical control.

Strategy is the case that tests the line, because directors approve it and the taxonomy places objective setting in management. Both are correct. Setting an AI objective, choosing the initiatives that pursue it, and executing them are management wherever they are performed. Approving the strategy those objectives serve, ratifying the risk tolerance the strategy must stay inside, and fixing the uses of AI the institution will not permit are oversight, because each settles the authority and the boundary within which management decides rather than the decision itself. A board that approves an AI strategy without ratifying the tolerance it runs inside has performed the management half of the act and left the oversight half to default.

One subject, three layers. The supply chain shows the taxonomy at work, because it appears at every layer with a different question at each. At the oversight layer: who accepted this vendor's residual risk, by name; whether the vendor sits inside the program's scope or has been declared out, and who accepted that; whether independent assurance has been obtained over the vendor, on what scope, as of what date. At the management layer: vendor selection and due diligence, contract terms and the right to audit, notification duties, ongoing monitoring, offboarding. At the technical layer: the provenance and integrity of the models, data, and components the system is built from, and what a third-party model is permitted to reach. No layer answers the questions of the layer above or below it. That is what it means for layer assignment to follow subject matter rather than altitude.

The oversight program at a glance

Charter the program, then operate it. Elements one, two, and five are the minimum viable form for a resource-constrained institution.

1 Charter the program: its mandate, its scope, its granularity, and its required outputs

One page, amended onto an instrument the organization already keeps. Records which external instruments the institution has elected to apply and at what depth, which elements apply in full, which minimally, which are deferred, and who decided. A voluntary standard becomes a commitment the moment someone elects to apply it, and it is the commitment, not the standard, that assurance later tests. Everything below runs from this.

2 Assign and record named accountability, at two levels

Names the committee that holds oversight of the program by charter and the officer who answers to it for the program by name. Uses the sector's existing rule for naming an accountable officer where one exists, and rests on the governing body's own authority where it does not. Each officer who owns a domain that AI reaches holds a monitoring system within it and a recorded duty to escalate out of it. Records residual risk acceptances and out of scope declarations, each with a named acceptor and a review date. An out of scope declaration with no named acceptor and no expiry is indistinguishable from an omission, and a program that names the board and no one beneath it has answered half the question.

3 Set and ratify AI risk tolerance at the accountable level

Drift thresholds, safety margins, and operating limits are otherwise set locally, as engineering defaults, with no one recorded as having ratified any of them. Ratification converts a calibration into a decision the institution can stand behind.

4 Grant, bound, and periodically re-ratify authority to operate

Dated, scoped, expiring. Treats operation without real time human review as a separate grant of authority rather than an engineering consequence, because the decision to remove the human from the loop is a decision, not a design outcome. Authority that does not expire is not authority, it is inertia.

5 Establish the reporting, escalation, and disclosure line and its cadence

Routes the evidence the lower layers produce to a party who can act on it, on a standing cadence that does not depend on events. The report must support a decision to continue, constrain, or withdraw, and it must carry the adverse information, incidents, findings, inquiries, and deviations from tolerance, not only progress. Fixes escalation triggers and materiality thresholds before deployment, and gives at least one pathway to the governing body that does not pass through management's discretion. Fixes the disclosure position toward regulators and the market before an event, including any clock that statute attaches to it. Requires that the response to any escalated matter be recorded: what was asked beyond management's assurance, what was done, and when it was resolved. An institution that has not fixed its criteria in advance will fix them under pressure, and a red flag received without a recorded response is, on the record, a red flag ignored.

6 Establish independent assurance over the oversight program

Every control beneath the oversight layer is performed by the functions that manage, build, or operate AI. The object of independent review is the commitment, not the standard: did the organization apply what it said it would, at the depth it recorded, with the deferrals it declared.

7 Establish competency at the accountable level

The accountable officer needs enough to interrogate a report, not to validate a model. Competency is what makes the reporting line an instrument of oversight rather than a ritual of receipt.

Proportional, and not new

NOT ONE SIZE

Elements one, two, and five carry most of the value for a small institution, with element four added wherever any system acts without real time human review. A rural water system and a clearing bank perform the same acts at very different depth, and the charter makes that depth a recorded decision rather than a default.

This is why an oversight instrument scales from a small operator to a global group without a separate small entity version.

NOT INVENTED

Every element is drawn from existing authority: the GOVERN function of the NIST AI Risk Management Framework; ISO/IEC 38507:2022, which addresses the governance implications of the use of AI by organizations and is written for members of a governing body; the named designation mechanisms already mandatory in a majority of regulated sectors; and NIST's own addition of GOVERN to Cybersecurity Framework 2.0.

Not a rewrite, and not a new committee.

WHAT THIS MEANS FOR YOUR BOARD

- 1 Can we produce the document that records which external instruments we elected to apply, at what depth, and who approved that election?
- 2 Can we name the committee that holds oversight of the program by charter, the officer who answers to it for the program, and who accepted every system we have declared out of scope?
- 3 Can we show that our AI risk tolerance was set on a stated basis and ratified at our level, rather than inherited from engineering defaults?
- 4 For any system operating without real time human review, can we show a dated, scoped, expiring grant of authority, and the party with power to withdraw it?
- 5 Do escalation and disclosure reach us on triggers we fixed in advance, by at least one path that does not require management's permission?
- 6 If our minutes were produced tomorrow, would they show the report we received, the question we asked, the action we took, and an independent confirmation that the program operated as chartered?

Appendix: where each element sits

The seven elements are drawn from existing authority and correspond to structures already defined elsewhere. This table records the correspondence for readers who want to locate an element in the case law, the standards, or the fourteen governance domains of the Center's AI Oversight Program. It is a reference, not a requirement to adopt any of the instruments named.

ELEMENT	Q.	JUDICIAL AUTHORITY	STANDARDS AND REGULATION	PROGRAM DOMAIN
1 • Charter the program	i	<i>Caremark</i> (a reasonable reporting system); <i>Marchand</i> (name the mission critical risk)	NIST AI RMF GOVERN 1; ISO/IEC 42001 Clauses 4 and 5; ISO/IEC 38507	01 Governance Program and Policy Framework; 02 Governance Structure, Oversight, and Resources
2 • Named accountability, two levels	ii	<i>Boeing</i> (committee charter); <i>McDonald's</i> (officer duty within domain)	Sector designation rules; Regulation S-K Item 106(c) by analogy	02 Governance Structure, Oversight, and Resources
3 • Ratified tolerance	iii	<i>Marchand</i> (rigorous oversight of the named risk)	NIST AI RMF GOVERN 1.3 and MAP 1.5; ISO/IEC 42001 Clause 6.1	04 AI Risk Methodology, Scope, and Tolerance
4 • Bounded, expiring authority to operate	iv	<i>Caremark</i> second prong (monitoring a system once in operation)	ISO/IEC 38507 (delegation of decisions to AI); EU AI Act Article 14 (human oversight)	06 Model Risk and Agentic Lifecycle Oversight; 08 Transparency, Explainability, and Human Oversight
5 • Reporting, escalation, and disclosure line	v	<i>Boeing</i> (agenda, unfiltered reporting, independent channel, documented response); <i>Clovis</i> (acting on red flags)	Regulation S-K Items 106(b) and 106(c); Form 8-K Item 1.05; state notification clocks	13 AI Risk Escalation and Disclosure Protocols
6 • Independent assurance over the program	vi	<i>Sorenson</i> (documented, monitored program earns dismissal)	NIST AI RMF MEASURE 1.3; ISO/IEC 42001 Clause 9.2; the IIA Three Lines Model	03 Governance Program Assurance and Continuous Learning; 14 Validation of Escalation and Governance Effectiveness
7 • Competency at the accountable level	vi	<i>Boeing</i> (board unable to interrogate what it received)	Regulation S-K Item 106(c)(2) (management expertise); ISO/IEC 38507 (governing body competence)	02 Governance Structure, Oversight, and Resources

Cases cited. *In re Caremark Int'l Inc. Derivative Litig.*, 698 A.2d 959 (Del. Ch. 1996); *Marchand v. Barnhill*, 212 A.3d 805 (Del. 2019); *In re Clovis Oncology, Inc. Derivative Litig.*, C.A. No. 2017-0222-JRS (Del. Ch. Oct. 1, 2019); *In re The Boeing Co. Derivative Litig.*, C.A. No. 2019-0907-MTZ (Del. Ch. Sept. 7, 2021); *Firemen's Ret. Sys. of St. Louis v. Sorenson*, C.A. No. 2019-0965-LWW (Del. Ch. Oct. 5, 2021); *In re McDonald's Corp. Stockholder Derivative Litig.*, 289 A.3d 343 (Del. Ch. 2023).

Relation to Center Paper No. 1. The Caremark Liability Roadmap sets out six board-level actions and four officer-level actions that describe the same program from the board's seat. Its committee designation corresponds to element 2; its standing reporting, escalation protocols, and substantive reporting to element 5; its documentation in minutes to the record in question vi; its board competency to element 7; and its officer-level actions to element 2. Elements 1, 3, and 4, the election of instruments at a recorded depth, ratified tolerance, and a bounded, expiring authority to operate, are where this paper goes beyond the Roadmap.



The Center for AI Oversight is a 501(c)(3) nonprofit educational institution defining oversight governance for AI in regulated industries. Center materials are educational, and the Center takes no positions on matters involving specific institutions.

cfaio.org · info@cfaio.org

CENTER PAPER NO. 7 • SEPTEMBER 2026