Can an AI Policy Really Protect Your Organization?

Why AI Governance Must Be Designed, Not Merely Declared

Maziyar Bahrami | August 6, 2026 | Research Essay

The approval of an AI policy can feel like the moment an organization has taken control.

The document may prohibit employees from entering confidential information into public AI systems. It may require human oversight for consequential decisions, restrict unapproved tools, assign responsibility to an AI committee, and express commitments to fairness, transparency, privacy, security, and accountability.

These are reasonable provisions. Some are essential.

But approving them does not create the capabilities they assume.

The policy does not reveal that a marketing team has begun analysing customer feedback through an external language model. It does not detect that an AI feature has been added to software approved two years earlier. It does not know that a recruiter is gradually treating a vendor's ranking as the default basis for screening candidates. It cannot determine whether the person nominally "in the loop" has enough information, time, or authority to challenge the system.

Nor can the document prevent an AI agent from accessing the wrong records, reconstruct a decision that was never logged, or suspend a system whose behaviour has changed since approval.

The policy may be sound. The organization may still lack control.

This distinction is the starting point for a more serious understanding of AI governance. A policy defines what an organization intends to permit, prohibit, and protect. Governance must make those intentions visible, actionable, enforceable, and revisable in everyday work.

That is why AI governance is not only a policy or compliance problem. It is also a **sociotechnical design problem**.

The difficult question is not simply whether the organization has declared responsible principles. It is whether those principles have been designed into the roles, workflows, interfaces, permissions, technical systems, monitoring arrangements, and decision processes through which AI actually operates.

The problem is not a shortage of principles

Organizations are not starting from an ethical vacuum.

In their analysis of 84 AI ethics guidelines, Jobin, Ienca, and Vayena (2019) found considerable convergence around principles such as transparency, fairness, responsibility, privacy, and the prevention of harm. But they also found substantial differences in how those principles were interpreted, prioritized, and connected to implementation.

That difference matters.

Agreement that AI should be fair does not establish which harms should be examined, which affected groups are relevant, which measures should be used, or what level of disparity should prevent deployment.

A commitment to transparency does not specify what must be explained, to whom, at what stage, or in what form.

A requirement for human oversight does not identify the responsible person, the evidence that person should receive, the decisions they may override, or the consequences of raising an objection.

The language of principles can create broad consensus precisely because it leaves many difficult choices unresolved.

Mittelstadt (2019) argues that principles alone cannot guarantee ethical AI because abstract values remain open to competing interpretations and may have little practical force without professional duties, institutional structures, enforcement, and avenues for redress. Morley et al. (2020) describe the same implementation challenge as the movement from the **what** of AI ethics to the **how**: from naming desirable values to developing methods through which those values influence data, system requirements, development, deployment, and evaluation.

This is where many organizational efforts become weak. They concentrate on improving the statement of intent while leaving the operating environment largely unchanged.

The central problem is therefore not necessarily that the policy contains the wrong values.

It is that the values remain detached from the systems through which AI is selected, configured, accessed, supervised, and changed.

The implementation gap is a design gap

Organizational AI governance is broader than rulemaking.

Mäntymäki et al. (2022) define organizational AI governance as a system of rules, practices, processes, and technological tools for ensuring that an organization’s use of AI remains aligned with its objectives, values, and legal obligations. Their analysis also places AI governance within an existing landscape of corporate governance, information technology governance, and data governance. AI is not governed in an institutional vacuum.

Papagiannidis, Mikalef, and Conboy (2025) develop this organizational view further. Their review conceptualizes responsible AI governance through three kinds of practice:

- **Structural practices**, which allocate authority, responsibility, and oversight;
- **Procedural practices**, which shape assessment, development, deployment, monitoring, and response;
- **Relational practices**, which coordinate the technical, managerial, legal, and social actors involved in the AI lifecycle.

A policy can inform all three. It cannot substitute for them.

The difference can be expressed simply:

> ***Policy defines the expected condition. Governance creates the organizational capacity to detect and correct the distance between expectation and reality.***

That capacity has to be designed.

Someone must decide what information flows into a governance decision. Someone must have the authority to approve, reject, restrict, or suspend a system. Workflows must specify where review occurs. Interfaces must present the information a reviewer needs. Technical architecture must enforce permissions and preserve evidence. Monitoring must detect when assumptions no longer hold.

Without these arrangements, a policy may communicate expectations but remain distant from the points at which AI creates real consequences.

What governance by design means

The expression **governance by design** should not be reduced to adding an ethics checklist to software development or encoding a few rules into a model.

In this essay, I use the term more precisely:

> ***Governance by design is the deliberate translation of organizational principles, responsibilities, and risk boundaries into the structures through which an AI system is developed, accessed, used, supervised, monitored, and changed.***

This includes technical design, but it extends beyond technology.

Governance by design involves:

- **organizational design:** who owns the system, who evaluates risk, who may approve it, and who can stop it;
- **workflow design:** when review occurs, what triggers escalation, and how exceptions are handled;
- **technical design:** permissions, data boundaries, logging, monitoring, release controls, and interruption mechanisms;
- **interaction design:** how uncertainty, limitations, evidence, and alternative options are presented to users;
- **incentive design:** whether employees are rewarded for responsible challenge or pressured to accept automated outputs;
- **learning design:** how incidents, complaints, overrides, and audit findings change future systems and policies.

A 2021 expert guidance note prepared at the request of the European Commission's Directorate-General for Research and Innovation offers a useful model through its concept of **Ethics by Design**. The note explicitly states that it is not formal EU guidance, but its conceptual structure is valuable. It moves from ethical principles to requirements, development guidelines, methodologies, and specific tools. The point is that abstract commitments must become increasingly concrete as they enter development and use.

The same document describes "accountability by design" in operational terms. It calls for mechanisms through which undesirable effects can be detected, stopped, and prevented from recurring; human actors can supervise and override systems; complaints can be raised and assessed; and relevant decisions and processes can be traced.

Governance by design extends this logic from the technical development process to the organization as a whole.

It asks not only whether an AI system possesses a desirable feature, but whether the surrounding organization has been configured to govern that feature in practice.

Designing visibility: the organization must first know what exists

An organization cannot govern AI it cannot identify.

This requirement is becoming harder to satisfy because AI no longer enters organizations only through large, clearly labelled technology projects. It arrives through browser tools, office applications, software updates, coding assistants, analytics products, customer-service platforms, vendor systems, and individual employee accounts.

An employee may not describe an embedded search function as an AI system. A department may purchase a service without knowing which external model processes its information. A vendor may add generative features after the original procurement review.

A policy can state that only approved AI may be used. But unless the organization maintains a sufficiently current view of its systems, models, agents, vendors, owners, data access, and purposes, the category of "approved AI" remains administrative rather than operational.

NIST's AI Risk Management Framework makes this distinction explicit. Its GOVERN function calls for mechanisms to inventory AI systems, clearly documented roles and responsibilities, ongoing monitoring, periodic review, and processes for safely decommissioning systems. Governance is described as a cross-cutting and continuing function, not as a document produced at the beginning of the process (NIST, 2023).

A useful AI inventory should therefore answer more than "Which tools have we bought?"

It should identify:

- the AI systems and agents in use;
- their intended organizational purposes;
- the processes and decisions they influence;
- the data, applications, and tools they can access;
- their internal and external owners;
- their model and vendor dependencies;
- the actions they can recommend or execute;
- the people or groups affected by their outputs;
- and the mechanisms through which access can be restricted or removed.

Visibility is not yet governance. But without visibility, governance begins from an incomplete representation of the organization.

The policy applies to the organization leaders believe they have. AI operates in the organization that actually exists.

Designing responsibility: a committee is not yet accountability

Many organizations respond to AI risk by creating a committee.

A cross-functional group may bring together legal, compliance, security, data, technology, human resources, and business representatives. This can be valuable because AI risks rarely fit within a single organizational function.

But assembling relevant people does not automatically produce accountable decisions.

The committee may receive incomplete information. Its members may review only the part of the system that falls within their existing mandate. Recommendations may be advisory. No participant may possess clear authority to delay deployment. When several functions approve their respective components, responsibility for the combined outcome can become unclear.

A governance body becomes operational only when its decision rights are designed.

That requires answers to questions such as:

- Which systems must reach the committee?
- What evidence must be provided?
- Which risks can be accepted, and by whom?
- What conditions may be attached to approval?
- Who may require redesign or additional testing?
- Who may suspend the system after deployment?
- How are disagreements resolved?
- Who remains responsible when several functions share control?

Structural governance practices matter because accountability requires more than the retrospective identification of someone to blame. It requires prospective clarity about who is expected to notice, decide, intervene, and learn.

Raji et al. (2020), for example, propose an end-to-end internal auditing process in which organizational principles feed into evidence and review throughout the development lifecycle. Their framework connects accountability to assigned roles, interdisciplinary evaluation, documented decisions, and the capacity to abandon a system when its risks outweigh its benefits.

The question is not merely whether an organization has an AI ethics board.

It is whether that board, or another designated authority, receives the right signals and possesses the power to change what happens next.

Designing human oversight: presence is not control

One of the most common statements in AI policies is that a human must remain “in the loop.”

The phrase is reassuring because it appears to preserve human responsibility. But it often conceals more questions than it answers.

Which human?

At what point in the process?

Reviewing what information?

Under what time constraints?

With what knowledge of the system?

Can the person reject the output, request further evidence, delay the process, or suspend the system?

A person who approves hundreds of recommendations, sees only the model's preferred answer, and is evaluated primarily on speed may formally participate in the process while exercising little independent judgment.

Human oversight is therefore not created by placing a person at the end of an automated workflow. It must be designed as a meaningful decision capability.

Raisch and Krakowski (2021) show why automation and human augmentation cannot be separated cleanly. When AI automates part of a task, it also changes the remaining human work. It can alter attention, expertise, discretion, workload, and responsibility.

Kellogg, Valentine, and Christin (2020) similarly show how algorithmic systems reshape organizational control by recording, rating, recommending, restricting, rewarding, directing, and replacing aspects of work. The relevant object of governance is therefore not only the model. It is the configuration of people, technologies, incentives, and tasks through which an outcome is produced.

Meaningful human oversight requires at least four designed conditions.

First, the reviewer must receive relevant information. A model output presented as an authoritative answer creates a different form of oversight from an interface that displays uncertainty, missing information, limitations, and alternative evidence.

Second, the reviewer must possess appropriate competence. Training should address not only how to use the system, but when its outputs should be doubted.

Third, the reviewer must possess genuine authority. The right to object is meaningless if organizational targets, interface design, or managerial pressure make objection practically impossible.

Fourth, the decision must be traceable. The organization should be able to reconstruct what the system recommended, what the reviewer knew, what action was taken, and why.

The EU AI Act reflects this design logic for high-risk AI systems. Article 14 requires such systems to be designed and developed with appropriate human–machine interface tools so that they can be effectively overseen. It also states that oversight measures should be proportionate to the system's risk, autonomy, and context of use. The precise legal obligation depends on the system and the organization's role, but the underlying principle is clear: human oversight must be made technically and organizationally possible.

The presence of a human is an organizational fact.

The capacity to exercise judgment is a design achievement.

Designing architecture: policy must alter what the system can do

Consider a policy that prohibits confidential information from being entered into unauthorized AI services.

The rule may be communicated through training and employee guidance. But the document cannot identify sensitive information in a prompt, restrict access to an unapproved provider, preserve an audit trail, or alert a responsible control owner.

Those outcomes require mechanisms such as:

- approved AI environments;
- identity and access management;
- data classification;
- AI gateways;
- data-loss-prevention controls;
- logging and observability;
- procurement restrictions;
- release gates;
- incident-response processes;
- and clearly assigned control owners.

The same distinction applies to model evaluation. A policy can require testing before deployment. Unless the development and release process prevents an untested system from moving into production, testing remains an expectation rather than a constraint.

Architecture is where some organizational intentions become enforceable.

This does not mean that every ethical or governance judgment can be automated. Technical controls embody assumptions and cannot resolve every conflict involving fairness, human rights, uncertainty, organizational purpose, or acceptable risk.

But the opposite position is also inadequate. Human responsibility without supporting architecture leaves people accountable for systems they may be unable to observe, reconstruct, constrain, or stop.

Agentic AI makes this problem more visible.

An AI agent may be able to search documents, retrieve personal data, call software tools, update records, communicate with other systems, or initiate transactions. A policy saying that the agent should use “only the information necessary for its task” does not define an authorization system.

The design must specify:

- the identity under which the agent operates;
- the resources it may access;
- the actions it may perform;
- the context in which permission applies;
- whether access is temporary or persistent;
- which actions require human approval;

- which actions are reversible;
- what the system records;
- and how authority can be withdrawn.

For a simple workflow, these permissions may be largely predetermined. For a more autonomous agent, permissions may depend on changing relationships among the user, the task, the organizational role, the data, and the requested action.

The governance question therefore moves from a broad statement such as “agents must operate responsibly” to an architectural question:

> *Under exactly what conditions can this agent observe, decide, communicate, and act?*

Without that translation, the organization has a rule about autonomy but no reliable mechanism for governing it.

Designing for actual work

Governance often originates centrally. AI-supported decisions occur throughout the organization.

That creates a tension between consistency and context. The organization needs shared principles and minimum controls, but the meaning of those controls depends on the work being performed.

Using a language model to improve the wording of a public announcement is not equivalent to using one to summarize confidential research, screen job candidates, advise a clinician, or prepare a legal argument. The same general technology can create very different risks because the affected people, acceptable errors, evidentiary requirements, and possibilities for redress differ.

Selbst et al. (2019) describe several “abstraction traps” that arise when algorithmic fairness is treated as a property of a technical model detached from its social and institutional context. A technically coherent intervention may fail because the problem was framed too narrowly or because a solution was transferred between settings without accounting for differences in institutions and practice.

Governance therefore has to enter the context in which the technology is used.

Rakova et al. (2021), drawing on interviews with responsible-AI practitioners, found that organizational structures, incentives, and accountability arrangements strongly affected whether responsible-AI initiatives could move from aspiration to practice. Practitioners frequently encountered unclear accountability, conflicting performance priorities, and reactive decision structures. More mature arrangements required responsible-AI work to be embedded across the product lifecycle rather than delegated to isolated individuals or specialist teams.

This is why AI governance must be connected to work design.

It must shape:

- procurement and vendor selection;
- use-case registration;
- data-access decisions;
- development and release processes;
- task allocation between humans and machines;
- performance targets;

- user interfaces;
- review and escalation procedures;
- incident reporting;
- post-deployment monitoring;
- and system retirement.

A policy that conflicts continuously with how work is organized will struggle to govern behaviour. Employees still face deadlines, productivity demands, resource limits, and local objectives.

Governance by design asks whether responsible conduct is practically supported at the point of work, rather than merely requested from a distance.

Designing for change: approval cannot be the end

Most written policies appear stable.

The systems they govern are not.

A provider may update a model. A team may connect it to a new dataset. Employees may use it for a purpose that was not included in the original assessment. A workflow may change. The affected population may change. An advisory system may gradually become the default basis for decisions.

Even when the model itself does not learn continuously, the sociotechnical arrangement around it evolves.

A system that was acceptable when approved can therefore become unacceptable later.

NIST addresses this through its four interconnected functions: **Govern, Map, Measure, and Manage**. GOVERN operates across the lifecycle, while the other functions establish context, evaluate risk, and support action. NIST also calls for post-deployment monitoring, user and stakeholder input, appeal and override mechanisms, incident response, recovery, change management, and the ability to deactivate systems whose performance becomes inconsistent with their intended use.

The EU AI Act similarly defines risk management for high-risk AI systems as a continuous and iterative process conducted throughout the system lifecycle. It links risk management to foreseeable misuse, post-market evidence, human oversight, quality management, and corrective or preventive action. Again, the exact obligations depend on legal classification and organizational role; the relevant point is that lifecycle governance cannot be reduced to the possession of an internal policy.

This lifecycle perspective changes the meaning of approval.

Approval should not mean:

> *We evaluated the system once and accepted it.*

It should mean:

> *Based on the present evidence, the system may operate under specified conditions, subject to continued observation and the authority to intervene.*

That is a much less reassuring statement.

It is also a more honest one.

The governance control loop

Governance by design explains what must be embedded in an organization.

A control-loop perspective explains how those arrangements should operate over time.

A functioning AI governance system repeatedly performs four connected activities.

1. Sense

The organization gathers evidence about what is happening.

It inventories systems and use cases, monitors performance, records incidents and complaints, tracks overrides, observes vendor and model changes, and identifies changes in law, organizational context, or affected populations.

The sensing question is:

> ***What is happening, and what has changed?***

2. Interpret

Relevant actors evaluate what the evidence means.

They examine technical performance, legal obligations, affected stakeholders, ethical tensions, uncertainty, and organizational risk tolerance. They determine whether the system remains consistent with its intended use and acceptable boundaries.

The interpretive question is:

> ***Does the system remain acceptable under current conditions?***

3. Act

The organization possesses mechanisms and authority to intervene.

It may require further testing, restrict access, alter a workflow, impose additional review, notify affected people, change a threshold, suspend the system, replace a vendor, or abandon the use case.

The action question is:

> ***Who can change or stop the system, and through which mechanism?***

4. Learn

The organization uses experience to change future behaviour.

Incidents, complaints, near misses, audits, overrides, and unexpected outcomes lead to revisions in policy, training, responsibilities, architecture, system design, procurement, and risk assumptions.

The learning question is:

> ***What must change because of what we have discovered?***

These activities must remain connected.

Monitoring without intervention creates reports, not control.

Authority without reliable information creates arbitrary decisions.

Intervention without learning allows the same weakness to return.

A policy that is never updated through operational experience gradually becomes detached from the organization it is supposed to govern.

The control loop is therefore not an alternative to policy. It is the operating system through which policy becomes adaptive governance.

From written commitment to designed governance

The difference becomes clearer when policy commitments are connected to design decisions, operating mechanisms, and evidence.

Written commitment	Design question	Operating mechanism	Evidence
Confidential information must not be shared with unauthorized AI systems.	Which data and services are restricted, and what exceptions are legitimate?	Approved environments, data classification, access restrictions, AI gateways, and escalation procedures.	Access records, alerts, exceptions, investigations, and control tests.
Consequential decisions require human oversight.	Which decisions require review, and what authority must the reviewer possess?	Review checkpoints, interface design, decision criteria, override rights, and escalation paths.	Review records, overrides, reasons, appeals, review times, and outcomes.
AI systems must be assessed for unfair effects.	Which groups, harms, measures, and thresholds are relevant to the use case?	Contextual risk assessment, disaggregated evaluation, periodic reassessment, and remediation.	Test results, accepted thresholds, exception decisions, remediation actions, and revalidation dates.
Only approved AI systems may be used.	What counts as an AI system, and how are embedded features and agents identified?	System inventory, use-case registration, procurement review, named ownership, and release controls.	Current inventory, approval records, owners, model versions, and deployment history.
Third-party AI must remain within organizational risk tolerance.	How will changes in the provider, model, contract, or functionality be detected?	Vendor assessment, contractual controls, change notifications, monitoring, and reassessment triggers.	Assessments, notifications, version records, incidents, and review decisions.
AI incidents must produce corrective action.	What constitutes an incident, who receives the signal, and who can intervene?	Reporting channels, severity criteria, investigation procedures, response authority, and corrective-action processes.	Incident records, response times, causal analysis, corrective actions, and follow-up testing.

The final column is especially important.

A policy demonstrates that the organization has stated a commitment.

Evidence helps determine whether that commitment has changed behaviour.

Documentation should not exist merely to produce a larger archive. It should allow the organization to reconstruct decisions, challenge assumptions, respond to affected people, detect recurring weaknesses, and determine whether its controls work under real conditions.

Governance can fail even when the policy was followed

There is a more difficult possibility.

An AI system may cause harm even though the policy was followed, the required forms were completed, and the correct committee approved the project.

The assessment may have examined the wrong risk. The model may have performed well while the surrounding workflow encouraged overreliance. Each function may have approved its own component while nobody examined the combined system. The vendor may have met its contractual obligations, although the organization deployed the product in a context the vendor never evaluated.

In such cases, additional reminders to follow the policy will not correct the underlying weakness.

The weakness lies in how the problem, system, roles, and controls were designed.

This is where compliance and control separate.

Compliance asks:

> ***Did we follow the established requirements?***

Control asks:

> ***Can we keep the system within acceptable boundaries as evidence, behaviour, and conditions change?***

An organization needs both questions.

But answering the first does not automatically answer the second.

When organizations treat policy approval as proof of governance, they risk producing a visible structure of responsibility without the operational capability required to exercise it. The policy becomes reassuring evidence that the issue has been addressed, while the most consequential decisions remain difficult to see, contest, or reverse.

That is the real danger of policy theatre.

What boards and senior leaders should ask

The question “Do we have an AI policy?” is no longer sufficient.

A more serious governance discussion begins with different questions.

What exists?

Can the organization identify its important AI systems, agents, use cases, owners, providers, data access, dependencies, and affected processes?

What is each system permitted to do?

Have general policy commitments been translated into permissions, risk boundaries, tests, monitoring thresholds, review points, and escalation rules?

Who can intervene?

Does a named person or function possess the information, competence, authority, and technical means to restrict, suspend, or redesign the system?

Is human oversight meaningful?

Can reviewers understand the system's role, see relevant limitations, challenge its output, and act without unreasonable organizational pressure?

What evidence is retained?

Can the organization reconstruct how an important AI-supported decision was reached, what evidence was available, who reviewed it, and what action followed?

What happens after failure?

Do incidents, complaints, overrides, and audit findings lead to changes in architecture, workflows, incentives, responsibilities, training, and policy?

These questions shift the discussion from the appearance of governance to the organization's actual capacity to govern.

Policies still matter

None of this makes written AI policies unnecessary.

Employees need clear boundaries. Technical teams need common requirements. Managers need a basis for allocating authority and resources. Regulators, auditors, customers, and affected individuals need to understand what the organization accepts responsibility for.

But policy must be placed in its proper role.

A serious policy establishes direction and constraints.

Governance by design translates them into organizational arrangements.

Architecture makes some requirements enforceable.

Work design preserves meaningful human judgment.

Monitoring creates evidence.

Authority enables intervention.

Learning changes the system when previous assumptions no longer hold.

The danger is not simply that an organization may violate its AI policy.

The deeper danger is that it may be unable to see whether the policy is working, unable to intervene when it is not, and unable to understand why it failed.

A written policy can declare responsibility.

Only a designed system of visibility, authority, human judgment, technical control, evidence, and learning can make that responsibility real.

References

Academic and institutional sources

European Commission, Directorate-General for Research and Innovation. (2021). *Ethics by Design and Ethics of Use Approaches for Artificial Intelligence* (Version 1.0).

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence. *Official Journal of the European Union*.

International Organization for Standardization. (2023). *ISO/IEC 42001:2023: Information technology—Artificial intelligence—Management system*.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*, 389–399. doi:10.1038/s42256-019-0088-2.

Kellogg, K. C., Valentine, M. A., & Christin, A. (2020). Algorithms at work: The new contested terrain of control. *Academy of Management Annals, 14*(1), 366–410. doi:10.5465/annals.2018.0174.

Madaio, M. A., Stark, L., Vaughan, J. W., & Wallach, H. (2020). Co-designing checklists to understand organizational challenges and opportunities around fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. doi:10.1145/3313831.3376445.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics, 2*, 603–609. doi:10.1007/s43681-022-00143-x.

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence, 1*, 501–507. doi:10.1038/s42256-019-0114-4.

Mökander, J., Morley, J., Taddeo, M., & Floridi, L. (2021). Ethics-based auditing of automated decision-making systems: Nature, scope, and limitations. *Science and Engineering Ethics, 27*, Article 44. doi:10.1007/s11948-021-00319-4.

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. *Science and Engineering Ethics, 26*, 2141–2168. doi:10.1007/s11948-019-00165-5.

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. doi:10.6028/NIST.AI.100-1.

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible artificial intelligence governance: A review and research framework. *The Journal of Strategic Information Systems, 34*(2), Article 101885. doi:10.1016/j.jsis.2024.101885.

Raisch, S., & Krakowski, S. (2021). Artificial intelligence and management: The automation–augmentation paradox. *Academy of Management Review, 46*(1), 192–210. doi:10.5465/amr.2018.0072.

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. doi:10.1145/3351095.3372873.

Rakova, B., Yang, J., Cramer, H., & Chowdhury, R. (2021). Where responsible AI meets reality: Practitioner perspectives on enablers for shifting organizational practices. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), Article 7. doi:10.1145/3449081.

Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68. doi:10.1145/3287560.3287598.

Practitioner sources consulted

Ayres, B. (2026, July 27). *Why written AI policies alone won't protect your organization*. Teneo.

Cybermaniacs. (2026, February 12). *Why AI governance fails without human–AI work design*.

Ferretti, S. (2025, December 13). *Why policies alone, in their current form, will not save us*. LinkedIn.

Gilot, P. (2026, June 11). *AI governance is not a compliance problem. It is an architecture problem*. LinkedIn.

Kavanagh, J. (n.d.). *The design gap in AI governance*. AI Career Pro.

Meskarian, M. (2026, January 1). *Why AI governance efforts fail*. Medium.