AI Governance as Organizational Design

A Decision-Rights Framework for Human-Agent Enterprises

Pierre E. Neis
Menschgeist · The AO Method
agile-organization.com
August 2026

ABSTRACT

Enterprise AI governance is converging on a management-based regulatory paradigm — internal processes mandated by frameworks such as the EU AI Act, the NIST AI Risk Management Framework, and ISO management-system standards — yet adoption data show this paradigm producing broad but shallow compliance: 85 percent of organizations report implementing responsible-AI practices, while only 25 percent report a fully mature framework, and internal actors, not external regulators, are named as the primary accountability holders in the large majority of cases [G5]. This article argues that the shortfall is not a maturity gap that more policy will close, but a category error: AI governance has been treated as an extension of IT risk and compliance management rather than as a problem of organizational design — specifically, of decision rights, autonomy calibration, semantic alignment, coupling architecture, strategic cadence, and engineering discipline. Drawing on the Adaptive Organization (AO) Method, a systemic framework originally developed to diagnose and redesign human decision rights, autonomy, and organizational structure, this article develops and grounds six governance moves in the current academic and industry literature on human-AI task allocation, agentic-AI risk tiering, enterprise ontology and knowledge-graph engineering, sociotechnical coupling theory, adaptive strategy deployment, and software-engineering discipline as a control layer for AI-agent-authored work. Each move is shown to have a direct, independently developed analogue in the technical and management literature, suggesting convergent recognition — from regulatory scholarship, safety science, and software engineering alike — that agent governance is inseparable from the organizational structures agents are embedded in. The article concludes with a synthesis map translating AO's existing organizational-design vocabulary into AI governance practice, and discusses the framework's limitations and the empirical work still needed to validate it as an intervention rather than a well-evidenced diagnosis.

Keywords: AI governance, agentic AI, decision rights, organizational design, autonomy calibration, ontology, sociotechnical systems, Hoshin Kanri, domain-driven design, management-based regulation

1. Introduction

1.1 The governance gap adoption data reveal

By 2025, a majority of enterprises had moved past AI experimentation into some form of agentic deployment: 62 percent of organizations report at least experimenting with AI agents, and 23 percent report scaling an agentic system somewhere in the enterprise [2.9]. Yet the same research documents a widening gap between deployment and governance. Respondents report acting to manage an average of four AI-related risks in 2025, up from two in 2022 — but 51 percent of organizations using AI report having experienced at least one negative consequence, with nearly a third citing AI inaccuracy specifically [2.9]. Separately, Gartner forecasts that more than 40 percent of agentic AI projects will be cancelled by the end of 2027, attributing the failure not to model capability but to "escalating costs, unclear business value, or inadequate risk controls" [2.8]. The most detailed available survey of enterprise responsible-AI (RAI) practice quantifies the shape of the gap precisely: the share of organizations implementing RAI rose from 52 percent in 2022 to 85 percent in 2025, but the share reporting a fully mature RAI framework rose only from roughly flat to 25 percent over the same period, and nearly a third of RAI initiatives are less than a year old [G5]. The same survey finds that 87 percent of organizations name internal entities — not external regulators — as primarily holding them accountable for AI use, versus 7.5 percent naming regulators [G5]. Read together, these figures describe organizations that have adopted the language of AI governance faster than they have built the organizational structure that would make governance effective — and that, by their own admission, understand the accountability problem to be theirs to solve internally, not a matter of waiting for external rules.

This article treats that gap as diagnostic. If governance were primarily a regulatory-compliance problem, broader regulatory coverage (the EU AI Act entered into force in 2024; the NIST AI RMF was published in 2023) should correlate with narrower implementation gaps. Instead, the gap is opening at the level regulation does not reach: the internal allocation of decision rights, the calibration of autonomy to organizational readiness, the resolution of semantic disagreement between systems, the pacing of human oversight against agent execution speed, the cadence by which governance itself adapts, and the engineering disciplines that make agent behavior legible without requiring a human to review every action. These are organizational-design problems, and this article's central claim is that they should be approached with organizational-design methods.

1.2 Two competing paradigms for AI governance

The literature on AI governance for businesses can be read as split between two paradigms, both legitimate but pointed at different failure modes.

The first — call it the compliance-extension paradigm — defines AI governance as an extension of existing IT and data governance: "AI governance regulates the exercise of

authority and control over the management of AI," and can be decomposed "into governance of data, (ML) models and (AI) systems along four dimensions," relating directly to "existing IT and data governance frameworks and practices" [G2]. This paradigm has produced most of what currently exists in practice: approval checklists, model cards, risk registers, and the management-based regulatory instruments — the EU AI Act, the NIST AI RMF, and ISO/IEC 42001 among them — that Coglianese and Crum identify as sharing "a core management-based paradigm" that "seek[s] to motivate an increase in human oversight of how AI tools are trained and developed" [G3]. Coglianese and Crum's diagnosis is important: management-based regulation asks organizations to build internal processes, but "refinements and systematization of human-guided training techniques will thus be needed to fit within this emerging era" [G3] — that is, the regulation names the destination without specifying the organizational route.

The second — the organizational-design paradigm — treats AI's arrival as "an organization-design problem" in which firms must "allocate decision rights, define accountability across the AI lifecycle, and specify the evidentiary and procedural conditions under which authority can be safely delegated to AI-based systems" [1]. Mäntymäki et al.'s hourglass model of organizational AI governance makes the same move explicitly at the level of theory, noting that while "a multitude of AI ethics principles have been proposed to tackle these risks… the outlines of organizational processes and practices for ensuring socially responsible AI development are in a nascent state" [G1]. Dobbe goes further, arguing from a system-safety perspective that the dominant technical framing of AI safety — visible in reports produced by dozens of internationally nominated experts — itself "frustrates meaningful discourse and policy efforts to address safety comprehensively," and that the corrective is not to add non-technical factors on top of technical ones but to integrate them, because "the system safety discipline has dealt with the safety risks of software-based systems for many decades, and understands safety risks in AI systems as sociotechnical" [4.3, G3].

This article positions itself inside the second paradigm without rejecting the first. Management-based regulation is a necessary scaffold — it establishes that internal process matters and creates external pressure to build it — but it is deliberately silent on which internal design choices satisfy its requirements, and reasonably so: regulation that specified organizational structure in detail would be brittle across industries and firm sizes. The gap it leaves is exactly where organizational-design methods are needed, and exactly where this article's six governance moves are aimed.

1.3 The AO Method as a source of governance primitives

The Adaptive Organization (AO) Method is a systemic diagnostic and redesign framework developed for organizational agility: it maps decision flow, calibrates team autonomy against risk, diagnoses informal power and candor gaps, and integrates strategy deployment (Hoshin Kanri) into a living operating cadence rather than a static annual plan. None of these tools were built for AI. This article's argument is that they transfer with minimal modification, because the underlying problem — who decides what, under what constraints, reviewed how often, against what shared definitions — does not change when one of the decision-making agents is a machine rather than a person. What changes is the speed, the volume, and the fact that agents inherit organizational ambiguity rather than negotiating around it the way experienced employees learn to.

Six moves make this transfer concrete, corresponding to sections 2 through 7 below: (1) partitioning decisions by ambiguity rather than by task type; (2) reusing an existing autonomy-risk boundary — reversibility, blast radius, correlation — to calibrate agent autonomy; (3) resolving ontological disagreement before scaling agents that depend on contested definitions; (4) governing the pace of coupling between agent execution speed and human review cadence, rather than treating substitution as the operative question; (5) running governance as a living Hoshin Kanri spine rather than a static charter; and (6) using software-engineering discipline — Extreme Programming, Feature-Driven Development, Domain-Driven Design — as a structural governance layer for agent-authored code and workflows. Section 8 synthesizes the six moves into a single translation map, and Section 9 discusses the framework's limitations and the evidence still needed to validate it.

2. Move One — Partition by Ambiguity, Not by Task

2.1 The problem with task-based partitioning

The most common first question organizations ask about AI agents is “what can it do?” — a capability question that leads naturally to task-based partitioning: agents get assigned to task categories (drafting, data entry, scheduling) and humans retain everything else by default. The AO Method's first governance move rejects this framing. The relevant partition is not between task types but between decisions where meaning-making must stay human and decisions where execution is explicit enough to hand to an agent — a distinction that cuts across, rather than along, conventional task boundaries. A single customer-service interaction, for instance, might contain both kinds of decision: routing and status lookup are structurally explicit; whether to make an exception to policy for a distressed customer is not.

2.2 Literature support: delegability as a multi-factor, preference-sensitive judgment

The empirical study of task delegability supports treating delegation as a judgment about ambiguity and trust rather than a technical capability question. Lubars and Tan's foundational empirical treatment tested a purpose-built dataset of 100 tasks against four delegability factors — motivation, difficulty, risk, and trust — and found that “people do not want full automation even where it is technically feasible”: there is “little preference for full AI control and a strong preference for machine-in-the-loop designs, in which humans play the leading role,” with trust the factor most correlated with preferred delegation configuration [1.2]. This is a preference-level finding, but it converges with the AO argument at the level of mechanism: task difficulty (a capability variable) is not what drives people's delegation preferences; something closer to trust in outcome legitimacy — a proxy for ambiguity and accountability — is.

Sun's 2026 framework for authority and responsibility allocation in human-AI collaborative decision-making is the closest published analogue to the AO partition move, and is worth treating as a primary theoretical anchor for this section. Sun explicitly frames AI adoption as creating “an organization-design problem” and decomposes authority into seven decision rights — information access, recommendation, selection, approval, veto, execution, and escalation — mapped against five responsibility domains: system design, decision process, outcome stewardship, oversight, and remediation [1]. The framework's central mechanism, rights-control-responsibility alignment, states that delegating a decision right shifts effective control and evidence access, and that governance fails specifically when “the responsible actor lacks the competence, authority, or information needed to intervene” [1]. This is a formalization of exactly the AO intuition: the partition question is not “is this task hard,” but “does the party who will be accountable for this decision have what they need to actually exercise judgment over it.” Sun's own validation — two worked applications claiming the framework “produces more precise governance than a generic human-in-the-loop

requirement” [1] — supports treating decision-right decomposition, rather than blanket human-in-the-loop mandates, as the more defensible design unit.

Notably, delegation direction is not fixed to run from human to AI. Hemmer et al.'s controlled experiment with 196 participants reversed the usual framing, giving the AI agent the right to decide which task instances to hand to a human. Both task performance and task satisfaction improved under AI delegation, and — critically — the improvement held “regardless of whether humans are aware of the delegation,” with increased human self-efficacy identified as the mediating mechanism [1.3]. This finding matters for the partition argument because it demonstrates that decision rights are a genuinely bidirectional design variable, not an inherent human prerogative that AI governance merely needs to protect. The AO partition, in other words, is not a wall erected to defend human turf; it is a deliberately engineered allocation that in principle could be assigned either direction, and should be assigned according to where ambiguity and accountability actually sit.

2.3 Regulatory support: human oversight requirements presuppose a partition they do not specify

The EU AI Act's Article 14 (Human Oversight) requires that high-risk systems “be designed and developed in such a way… that they can be effectively overseen by natural persons during the period in which they are in use” [1.4], and Article 14(4) lists specific oversight capabilities regulators expect deployers to preserve — remaining aware of automation bias, correctly interpreting system output, deciding not to use the system in a given instance, and being able to intervene or halt it [1.4]. Article 14(5) goes further for certain systems, requiring a “four-eyes” verification by two competent persons before action is taken on system output [1.4]. Read as a checklist, these provisions specify what capabilities human overseers must retain; they do not specify which decisions require that oversight in the first place, or how an organization should draw that line. Enqvist's legal analysis of Article 14 makes this gap explicit and is, notably, a critique from within the regulatory-compliance literature itself rather than an outside objection: she finds that the article addresses “the functional rather than performative aspects of human control,” leaving “fairly sparse regulatory detail on the competency issues in human oversight,” which she calls “a risk and flaw of the Regulation” — one that risks “legitimising… a human oversight infrastructure that chiefly allows for mere superficial human involvement” [1.5]. NIST's AI RMF reaches a structurally similar conclusion from a different angle, explicitly assigning “role clarification” for human-AI configurations to the framework's Govern function rather than to a testing or measurement function [1.6] — placing the partition decision squarely in organizational governance, and away from technical assurance, exactly where the AO Method already operates.

NIST's own framework additionally flags the measurement gap this section is trying to close: “the degree to which humans are empowered and incentivized to challenge AI system output requires further studies,” and badly organized human-AI teams “can amplify human biases, leading to more biased decisions than the AI or human alone,” while well-organized ones

“can result in complementarity and improved overall performance” [1.6]. The word doing the work in both sentences is organized — the same variable the AO partition move is designed to make explicit and revisable.

2.4 The AO partition in practice

Applying this literature to the AO partition produces a working rule: human-owned decisions are those where the accountable party's judgment cannot be reduced to a documented, checkable criterion without losing something regulators, customers, or the organization itself would recognize as material — legitimacy, ethical weight, or novel context the existing rules do not anticipate. Agent-owned decisions are those where goals, boundaries, and validation criteria are already explicit enough that a documented criterion is the decision, and where the residual risk from acting on it is something the organization has separately decided to accept (a question the autonomy-risk boundary in Section 3 makes explicit). This partition is not fixed: Sun's framework treats AI reliability as conditioning “the permissible scope of selection and execution rights” over time [1], meaning the AO partition should be a living map, revisited as agent capability and organizational trust in that capability change — not a one-time classification exercise.

3. Move Two — Reuse the Autonomy-Risk Boundary

3.1 From capability to permission

The AO Method already has a working calibration tool for autonomy, originally developed for team-level decision rights: reversibility, blast radius, and correlation. The second governance move is to apply the same lens to AI agents rather than build a parallel AI-specific risk-tiering framework from scratch. This section grounds that reuse in a body of 2025-2026 literature that has independently converged on the same core distinction — separating what an agent is technically capable of from what it is organizationally permitted to do.

Zheng et al.'s framework for agentic AI autonomy makes this separation the paper's organizing move: “discussions of autonomy often conflate what systems are technically capable of doing with what they should be permitted to do in practice” [2.1]. Their five-level structure — spanning reactive execution, decision support, supervised action, goal-directed autonomy, and delegated operational authority — separates Allowed Autonomy Levels (AAL), calibrated to risk, oversight, and accountability, from Autonomous Capability Levels (ACL), what the system can technically do [2.1]. Their worked case — an enterprise data-engineering agent assessed at high technical capability but deliberately constrained to a lower allowed autonomy level based on “risk, reversibility, and organizational readiness” [2.1] — is precisely the AO autonomy-risk calibration applied to a concrete agent deployment: the organization does not ask whether the agent could act autonomously, but whether it should be permitted to, given what would happen if it were wrong.

Feng, McDonald, and Zhang reach a compatible position from a different direction, framing autonomy explicitly as “a deliberate design decision, separate from its capability and operational environment,” and defining five escalating autonomy levels by the role a human occupies relative to the agent — operator, collaborator, consultant, approver, observer — rather than by what the agent does [2.2]. This role-based framing is a useful complement to the AO boundary: it clarifies that autonomy calibration is simultaneously a statement about the agent's permitted action and about what the accountable human is expected to be doing at each level, which maps directly onto the AO practice of pairing autonomy grants with an explicit escalation contract.

3.2 Reversibility and blast radius as the calibrating dimensions

The specific dimensions the AO boundary uses — reversibility and blast radius (correlated impact across systems) — are independently supported as the correct calibrating variables, rather than task category or model capability score. Mitchell et al.'s position paper on autonomous agents establishes the underlying monotonic relationship any calibration scheme must price in: “risks to people increase with the autonomy of a system: the more control a user cedes to an AI agent, the more risks to people arise,” with safety risk — risk to human life — flagged as “particularly concerning” [2.4]. This licenses treating autonomy itself, not merely the task an agent performs, as a risk-bearing variable that must be governed directly.

Engin and Hand's 2026 paper is the strongest available critique of static risk tiering and a direct ally of the AO Method's insistence that the boundary be a living map rather than a fixed table. They argue that categorical frameworks — “high-risk”/“low-risk,” fixed autonomy levels, human-in-the-loop/human-out-of-the-loop binaries — fail structurally because “risk is contextual, systems are continuously updated, and authority is relational in multi-agent settings” [2.3]. Their proposed alternative, dimensional governance, is built on “the 3As”: decision authority (“when and how should AI have the right to decide?”), process autonomy (“how independently does the system operate?”), and accountability configuration (“who is responsible when something goes wrong?”) [2.3]. Engin and Hand's framing of the core failure mode — “static categorisation assumes human-AI relationships can be neatly boxed into fixed states, when reality reveals continuous, evolving patterns of agency, autonomy, and accountability” [2.3] — is close to a direct restatement of why the AO Method treats autonomy boundaries as diagnostic outputs to be revisited, not compliance thresholds to be set once. Their further point that “monitoring system drift along dimensions becomes as important as assessing initial design intentions,” with thresholds acting as “governance pivots, triggering calibrated oversight responses” [2.3], anticipates the drift-detection argument this article returns to in Section 6. Notably, Engin and Hand are careful to position dimensional governance as complementary to, not a replacement for, categorical frameworks like the EU AI Act's risk tiers [2.3] — a stance this article adopts as well: the statutory tiers set a floor; the AO boundary calibrates autonomy above that floor at the operational level regulation does not reach.

The EU AI Act's own risk classification (Article 6) illustrates the categorical logic Engin and Hand critique, and is instructive precisely because its exemption language already gestures toward the AO Method's ambiguity/execution distinction without formalizing it. Article 6(3) exempts systems that perform “a narrow procedural task,” that “improve the result of a previously completed human activity,” or that assist a human assessment “without proper human review,” from high-risk classification — but reimposes high-risk status categorically wherever a system “performs profiling of natural persons,” regardless of those conditions [2.7]. The Act, in other words, already distinguishes structured-execution work from judgment work in its exemption logic; it simply does not give organizations a repeatable

internal method for making that same distinction across their own decision inventory, which is the gap the AO boundary is designed to fill.

3.3 Measuring autonomy in engineering artifacts, not only in behavior

A practical objection to any autonomy-calibration scheme is that behavioral autonomy is hard to observe directly, especially at scale. Cihon et al. address this by scoring an agent's orchestration code — the software that actually runs the agent — against a taxonomy of two autonomy attributes, impact and oversight, "eliminating the need to run an AI agent to perform specific tasks" [2.5]. This is directly useful for the AO Method's operational version of the boundary: autonomy is not only a behavioral property to be observed after deployment, but a property legible in engineering artifacts before deployment, which connects the autonomy-risk boundary (this section) to the software-engineering guardrails developed in Section 7.

3.4 Why the boundary matters commercially, not only ethically

The commercial case for calibrated rather than blanket autonomy is direct in the adoption data already cited: Gartner attributes the projected cancellation of more than 40 percent of agentic AI projects by 2027 to "escalating costs, unclear business value or inadequate risk controls," and specifically to organizations that let "hype" outrun a realistic assessment of "the real cost and complexity of deploying AI agents at scale" [2.8]. McKinsey's parallel 2025 survey of 1,993 respondents across 105 countries finds a similar pattern at the value-realization stage: "high performers are more likely than others… to say their organizations have defined processes to determine how and when model outputs need human validation to ensure accuracy," and running an agile, well-defined delivery organization is "strongly correlated with achieving value" from AI investment [2.9]. Both findings support the same conclusion as the academic literature from a different vantage point: undifferentiated autonomy, whether granted too generously or withheld too uniformly, is a commercial as well as a safety liability, and calibrating it against reversibility and blast radius is a defensible middle path supported independently by regulatory scholarship, safety research, and enterprise survey data.

4. Move Three — Resolve Ontology Before Scaling Agents

4.1 Ontology as governance schema, not documentation

An AI agent does not encounter an organization directly; it encounters the organization's systems, and inherits whatever definitions those systems encode — including definitions that conflict with each other. A governance document cannot repair a definition of “customer” that means three different things across three systems; it can only specify who was supposed to have caught that before an agent acted on it. This section grounds the AO Method's ontology-audit practice — pulling core entities from every system that touches them and naming an accountable owner for each contested definition — in the enterprise ontology and knowledge-graph literature, which treats semantic grounding as a control mechanism rather than a modeling nicety.

The clearest statement of this stance, and the anchor source for this section, is Tuan and Sanyal's three-layer ontological framework — Role, Domain, and Interaction ontologies — for grounding LLM-based enterprise agents [3.1]. Their controlled experiment (1,800 runs across five industries and three large language models) found that ontology-coupled agents “significantly outperform ungrounded agents on Metric Accuracy ($p < .001$) and Role Consistency ($p < .001$) across all three models with large effect sizes” [3.1]. Two further findings sharpen the argument. First, an inverse parametric knowledge effect: “ontological grounding value is inversely proportional to LLM training-data coverage of the domain” — meaning ontology matters most precisely in the specialized, enterprise-specific, or non-English domains where organizations most need reliable agent behavior and have the least ability to rely on a model's general training [3.1]. Second, a structural diagnosis with direct governance implications: “current enterprise systems constrain agent inputs (context assembly, tool discovery, governance thresholds) but not outputs” [3.1] — an asymmetric coupling in which organizations govern what an agent is given but not what it produces, which the authors propose closing with output-side ontological validation. The framework has already reached production scale, “serving 22 industry verticals with 650+ agents” [3.1], which supports treating ontology resolution as an operationally tractable governance move rather than a research aspiration.

4.2 Why the timing matters: agents query continuously, at machine speed

Enterprise governance playbooks reinforce why ontology resolution has to precede, not follow, agent scaling. Autonomous agents query enterprise data continuously and act on it at a pace occasional human review was never designed to supervise — meaning an unresolved definitional conflict does not surface once, on a predictable schedule, but propagates continuously across every agent decision that touches the contested entity. Laurenzi, Mathys, and Martin's account of enterprise knowledge-graph engineering explains why organizations postpone this work despite the stakes: “conventional knowledge engineering approaches… still require a high level of manual effort and high ontology expertise, which hinder their adoption across industries” [3.2]. Shimizu and Hitzler, writing an agenda-setting piece from within ontology engineering itself, argue that large language models can meaningfully accelerate “ontology modeling, extension, modification, population, alignment, as well as entity disambiguation,” but insist that “modular approaches to ontologies will be of central importance” [3.3] — a caution against treating LLM-assisted ontology work as a shortcut around the decomposition discipline the AO ontology audit already requires (per-entity ownership, side-by-side comparison of definitions across systems, explicit resolution before scaling).

Feng, Wu, and Meng's work on ontology-grounded knowledge-graph construction under the Wikidata schema demonstrates the specific mechanism by which grounding buys reliability: grounding generation in an authored ontology is what secures “consistency and interpretability,” not merely output accuracy, enabling “scalable KG construction… with minimal human intervention, that yields high quality and human-interpretable KGs” [3.4]. This distinction — consistency and interpretability as separate payoffs from accuracy — matters for governance because an agent can be individually accurate in any given interaction while still being organizationally inconsistent across interactions, precisely the failure mode ontology resolution targets. Qiang, Wang, and Taylor's Agent-OM system addresses the harder version of the same problem — reconciling two organizations' or two systems' different ontologies against each other, using LLM agents for retrieval and matching — and report results “very close to the long-standing best performance on simple… tasks” with “significant improve[ment]… on complex and few-shot” matching tasks [3.6], evidence that automated ontology reconciliation is becoming tractable at the technical layer even as the organizational decision of who owns a contested definition remains a human governance question.

4.3 The cost of not resolving ontology: quantified evidence

The commercial stakes of unresolved semantic inconsistency are unusually well quantified in recent industry survey data, which supports treating ontology resolution as material rather than aspirational. Semarchy's 2025 survey of 1,050 senior business leaders across the US, UK, and France found that while 74 percent of businesses planned to invest in AI initiatives, only 46 percent were confident in their data quality, and 98 percent "have suffered from AI-related data quality issues" [3.7]. The survey's named causes are directly ontology-shaped: duplicate records (25 percent) and inefficient data integration (21 percent) — both symptoms of the same entity meaning different things, or being represented inconsistently, across systems [3.7]. Semarchy's CTO frames the resulting risk in language that closely parallels this article's thesis: "deploying AI at scale on a bad foundation will only magnify business risks and lead to wasted investments" [3.7].

IBM's 2026 analysis, drawing on the IBM Institute for Business Value's 2025 CDO study, monetizes the same gap: 43 percent of chief operations officers name data quality as their most significant data priority, over a quarter of organizations estimate losses exceeding USD 5 million annually from poor data quality (7 percent report losses of USD 25 million or more), and data accuracy or bias concerns are cited as a leading barrier to scaling AI by nearly half — 45 percent — of business leaders [3.8]. The report's illustrative incidents are instructive for the coupling argument developed in Section 5 as well: Samsung Securities' 2018 invalid data entry, which "mistakenly trigger[ed] the issuance of billions of duplicate shares," was "identified within minutes" — fast detection did not prevent "market disruption, regulatory penalties, leadership resignations and an estimated hundreds of millions of dollars in market value loss" [3.8]. The lesson generalizes directly to agentic AI: detection speed alone does not substitute for resolving the underlying semantic or definitional fault before it can propagate, because propagation at machine speed can outrun even a fast human response.

4.4 The ontology audit, in practice

The AO Method's ontology audit operationalizes this literature into a repeatable practice: pull the same core entity — customer, deal, risk, "done" — from every system that touches it, and write each definition side by side; name an accountable owner for each contested term before scaling any agent that depends on it; apply the autonomy-risk lens from Section 3 to the ambiguity itself, resolving what has blast radius and deferring what does not; and audit live agent deployments specifically for inherited, unresolved definitions, since — as Tuan and Sanyal's production-scale evidence and Semarchy's 98-percent figure both indicate — this is a present, not a future, risk [3.1, 3.7].

5. Move Four — Govern Coupling Pace, Not Substitution

5.1 The wrong question and the right one

Organizations tend to frame AI adoption as a substitution question: does the agent replace the team, the role, the process? The AO Method's fourth governance move reframes the operative question as one of coupling pace: how tightly, and how fast, is the agent's execution loop connected to the human governance loop, and what happens when the two run at different speeds? This section grounds that reframing in sociotechnical systems theory, beginning with the classical accident-analysis literature and extending to contemporary agentic-AI management research that has arrived at compatible conclusions independently.

5.2 Coupling and complexity: the classical framework

Charles Perrow's Normal Accident Theory supplies the foundational vocabulary. As summarized by Pidgeon, Perrow's two core constructs are interactive complexity — “the number and degree of system interrelationships” — and tight coupling — “the degree to which initial failures can concatenate rapidly to bring down other parts of the system” [4.1]. Perrow's central claim, restated by Pidgeon, is that “when systems exhibit both high complexity and tight coupling... the risk of failure becomes high” [4.1]. Two further Perrow-derived points transfer directly to AI-agent governance. First, his counter-intuitive warning that safety devices can reduce safety if they add complexity without addressing coupling — illustrated by a 1972 aircraft accident in which “pilots could not diagnose the fault amid at least nine other cockpit warnings and alarms” [4.1] — is a caution against the reflexive governance response of adding more dashboards, alerts, and review gates to an agentic system without first addressing how tightly its actions are coupled to downstream consequences. Second, Perrow's loose-coupling contrast is clarifying: universities are “interactively complex but only loosely coupled,” while modern production lines are “tightly coupled” but structurally simple, and “neither... tends to suffer systemic accidents” [4.1] — the danger is specifically the combination, not either property alone. This maps onto agentic AI directly: an agent operating in a complex domain (many possible states, many stakeholders) that is also tightly coupled to irreversible action (no buffer, no review step, direct system writes) is the AO Method's high-risk quadrant from Section 3; complexity or coupling alone, without the other, is comparatively governable.

Macrae's re-analysis of the 2018 fatal Uber self-driving vehicle crash extends this classical framework specifically to autonomous and intelligent systems, deriving the SOTEC framework: five domains of sociotechnical risk — structural, organizational, technological, epistemic, and cultural [4.2]. Macrae's framing statement is a direct rebuttal to purely technical safety analysis: “AIS are necessarily developed and deployed within complex human, social, and organizational systems, but to date there has been little systematic examination of the sociotechnical sources of risk and failure in AIS” [4.2]. This is precisely the gap the AO Method's coupling-pace move is designed to close at the organizational level,

and it is worth noting that Macrae reaches this conclusion from safety-science methodology (constant comparative analysis of investigative reports) entirely independently of the organizational-design literature this article otherwise draws on — a second, methodologically distinct line of evidence converging on the same structural diagnosis.

5.3 Speed mismatch as a management problem, independently identified

The most directly usable contemporary evidence for the coupling-pace argument comes from Renieris et al.'s 2025 MIT Sloan Management Review / Boston Consulting Group study, based on a global executive survey (1,221 responses) and an international expert panel of more than 50 practitioners, academics, and policymakers. Sixty-nine percent of panel experts agreed that holding agentic AI accountable requires genuinely new management approaches, against 25 percent warning of "AI exceptionalism" — an unresolved but informative split [4.5]. The practitioner quotes gathered in this study function as direct, independently sourced testimony for the AO Method's coupling-pace thesis:

> "Today's workflows were not built with the speed and scale of AI in mind, so addressing gaps will require new governance models, clearer decision pathways, and redesigned processes that make it possible to trace, audit, and intervene in AI-driven decisions." — Shelley McKinley, GitHub [4.5]

> "Old management models, built for human-paced systems, fall short in tracking AI's dynamic behavior, risking unaddressed errors or harms." — David Hardoon, Standard Chartered Bank [4.5]

> "It's not enough to run a checklist once and call it done... to stay accountable, organizations need tools and processes that follow AI systems throughout their use." — Franziska Weindauer, CEO, TÜV AI.Lab [4.5]

The counter-position deserves inclusion for balance and because it sharpens, rather than undermines, the AO argument: Amit Shah of Instalily.ai argues that "agentic AI doesn't require a new management doctrine. It requires courage. We already govern systems that are fast, complex, and opaque like nuclear power, algorithmic trading, and aviation... True governance starts with a name, not a framework" [4.5]. This is not actually a rejection of coupling-pace governance — nuclear power and aviation are exactly the tightly-coupled, high-complexity domains Perrow's theory was built to analyze, and both rely on extensive engineered social brakes (checklists, dual authorization, black-box recorders, mandatory cool-down periods) rather than on courage alone. Ben Dias of IAG makes a related, more moderate point: "managers routinely delegate to team members whose decision-making processes they cannot fully predict or control, yet maintain accountability through clear boundaries, outcome-focused oversight, and appropriate monitoring" [4.5] — precisely the design pattern the AO Method's autonomy-risk boundary (Section 3) and coupling-pace move (this section) together implement for agents.

Kilian's complex-systems framework for AI risk supplies additional theoretical grounding for why coupling pace, specifically, generates risks distinct from accidents or misuse: rapid AI integration can produce “emergent, system-level dynamics beyond conventional, proximate threats,” organized into “antecedent structural causes, antecedent AI system causes, and deleterious feedback loops” that can “destabilize trust, shift power asymmetries, and erode decision-making agency across scales” [4.4]. Dobbe's system-safety critique, already introduced in Section 1, reinforces the same point about the need for architectural rather than additive intervention: the corrective to speed mismatch is not bolting a review step onto an existing fast loop, but redesigning the loop's structure so that human judgment enters at points the system's own tempo can actually accommodate [4.3].

5.4 Designing the social brake: HOTL, HITL, and management by exception

Where should the social brake sit? The Joint Air Power Competence Centre's analysis of rapid-response weapons systems — a domain with genuinely extreme coupling and stakes — supplies precise, transferable definitions: Human-In-the-Loop (HITL), where “the user has complete control to start or stop the automation”; Human-On-the-Loop (HOTL), giving the user “control only over autonomy planning”; and Human-Out-Of-the-Loop (HOOTL), where no human control point remains [4.6]. Two design concepts from this literature transfer directly to enterprise agent governance. First, the distinction between “management by consent” (active human approval required for each action) and “management by exception” (the human intervenes only when a deviation is flagged), recommended specifically “to mitigate automation bias by requiring the human operator to remain actively engaged, except in delta near-zero situations” [4.6] — this is the mechanism the AO Method's coupling-pace move formalizes as “social brakes”: sampling review and exception escalation, rather than either full manual gatekeeping or full autonomy. Second, the concept of graceful degradation: automation that fails suddenly and without warning is “brittle,” whereas well-designed systems ensure “human supervisors are informed and aware that automated features are compensating for performance deviations” [4.6] before a failure becomes visible only in its consequences. Applied to enterprise agents, this argues for designing observable, legible drift signals into agent workflows — a requirement that connects directly to the drift-detection practice in Section 6 and the logging findings in Section 7.

The practical implication for the AO Method is that coupling pace is not a single dial but a design surface with at least three variables: which position on the HITL/HOTL/HOOTL spectrum a given decision class occupies (itself derived from the autonomy-risk boundary in Section 3); whether the human checkpoint operates by consent or by exception; and how quickly a deviation becomes visible to the accountable human once it occurs. Fast technical systems paired with slow human review do not need slower agents — they need explicitly engineered brakes positioned at the right points in the loop, sized to the actual coupling and complexity of the decision, not applied uniformly across all agent activity.

6. Move Five — Governance as a Living Hoshin Spine

6.1 Why a static charter fails a system that keeps changing

AI governance documents are typically written once, approved, and revisited on an annual cycle if at all — a cadence borrowed from compliance and audit practice, where the underlying subject (a business process, a data-retention policy) changes slowly. Agentic AI systems do not hold still at that pace: new capabilities, new data sources, and shifting ontological boundaries can silently widen an agent's effective autonomy well beyond what was actually governed, long before the next scheduled review. The AO Method's fifth governance move borrows directly from its existing Hoshin Kanri integration — treating AI governance as a living strategic spine, with an annual guardrail layer setting direction (what agents are permitted to touch, what triggers escalation) and monthly rolling reviews adapting operational detail without losing overall coherence.

6.2 Hoshin Kanri and catchball as an alignment mechanism

Hoshin Kanri's canonical English-language treatment describes it as a methodology in which "the process of developing targets, the assessment of the means to achieve the targets and the deployment of both are fundamental to successful implementation" [5.1]. Its defining mechanism, catchball, is described in its original form as an idea "thrown around from person to person" like a children's ballgame — "a critical element requiring continuous communication to ensure the development of appropriate targets and means" at every level of the organization [5.1]. Ćirić et al.'s systematic review elaborates the mechanism further, noting that catchball (also called nemawashi) "encourages alignment, clarification, and employee involvement" through communication that is explicitly "two-way, top-down, bottom-up, and collaborative," with "every objective, plan, and problem... discussed at all levels" [5.2]. The purpose, in their formulation, is "to translate strategies into objectives at the lower levels, considering the relationship between causes and effects" [5.2] — which is exactly the function the AO Method asks catchball to serve for AI governance: translating an annual guardrail ("agents may not initiate customer-facing financial commitments without escalation") into the specific, evolving operational detail of what that means for a particular team's particular agent deployment this month, and feeding operational discoveries back upward when they reveal the annual guardrail itself needs revision.

Wilson, Cudney, and Marley's 2023 systematic review of Hoshin Kanri research (covering 1990-2021 publications) is honest about the method's evidentiary limits — noting that despite decades of Western use, "there is still relatively little published research on the topic," with most years producing fewer than two publications [5.3] — a limitation this article's own Section 9 returns to regarding the AO Method's Hoshin-based governance cadence specifically. Ćirić et al.'s review is similarly candid that most Hoshin Kanri case studies lack "concrete quantitative results" and "sufficient time horizons to establish whether Hoshin Kanri has long-term positive effects" [5.2]. This article treats Hoshin Kanri as a

well-established alignment mechanism with a plausible mapping onto AI governance cadence, while flagging — consistent with the source literature's own candor — that its long-run effectiveness even in its original domain remains under-evidenced, which counsels empirical caution rather than confident extrapolation.

6.3 Continuous governance in the 2026 literature

Independent support for treating AI governance as continuous rather than episodic comes from Lee et al.'s 2026 Co-Lifecycle Governance framework for learning medical AI, arguably the strongest available 2026 source on this specific question and a close structural analogue to the AO Hoshin spine. Lee et al. argue that “learning AI systems do not merely introduce new risks; they alter the temporal and organizational conditions under which responsibility is exercised,” and that consequently “oversight as a discrete event” must be replaced by “a continuous, synchronized process grounded in 4 pillars: continuous validation, agile change management, proactive performance surveillance, and distributed accountability” [5.4]. Their operational detail on continuous validation — “staged reassessment across 3 time horizons”: short (input drift, uncertainty, process indicators), intermediate (proxy outcomes, chart-review audits, surrogate markers), and long (confirmation after definitive outcomes mature) [5.4] — maps closely onto the AO Method's own layered cadence of annual guardrails and monthly rolling review, and supplies a more granulated time structure the AO Hoshin spine could adopt directly. Their insistence that “surveillance does not 're-prove' effectiveness; it detects deviations and triggers targeted validation when thresholds are crossed” [5.4] is a precise description of what the AO Method calls monthly rolling catchball functioning as a live deviation signal rather than a once-a-year cascade. Lee et al.'s closing claim — “learning AI will not wait for governance to catch up; oversight must evolve in lockstep with adaptive intelligence” [5.4] — is close to a direct restatement of this article's overall thesis, arrived at independently from the medical-AI regulatory literature.

Bakirtzis et al.'s proposal for dynamic certification supplies the regulatory-side counterpart to this operational argument, critiquing point-in-time conformity assessment on the grounds that “traditional regulatory mechanisms, often designed for static… technologies, find themselves at a crossroads when faced with the fluid and evolving nature of AI systems” [5.5]. Their proposed framework requires “common progress in multiple domains: technical, socio-governmental, and regulatory” and aims to keep systems aligned “bidirectionally with the real-world contexts in which they operate” [5.5] — a regulatory articulation of the same rolling, bidirectional cadence the AO Hoshin spine implements organizationally through catchball.

6.4 Bounded autonomy contracts and dual transparency

The AO Method's Hoshin-based governance also relies on bounded autonomy contracts: teams retain pivot rights inside an agreed impact envelope, while movement beyond that envelope triggers escalation — the same dual-transparency architecture the AO Method already uses to keep system-level signals visible to leaders without converting team-level learning spaces into compliance theatre. This design directly operationalizes Engin and Hand's point (Section 3) that “thresholds serve as governance pivots, triggering calibrated oversight responses” [2.3] and Lee et al.'s point that surveillance should trigger validation “when thresholds are crossed” [5.4] rather than requiring continuous manual review of every agent action. The practical benefit, consistent across both sources, is that governance attention is concentrated where drift has actually occurred rather than spread evenly and shallowly across all agent activity regardless of whether anything has changed.

7. Move Six — Guardrails Built Into the Work

7.1 From governance-by-inspection to governance-by-constraint

The preceding five moves govern decisions: who owns them, how autonomous the agent making them may be, whether the concepts underneath them are resolved, how fast the human loop reacts, and how often the whole system recalibrates. This section addresses a different governable surface: agents that write or execute code and workflows directly. Here, the AO Method's sixth move argues that established software-engineering disciplines — Extreme Programming's test-first practice, Feature-Driven Development's bounded work interfaces, and Domain-Driven Design's ubiquitous language and bounded contexts — function as governance infrastructure in their own right, constraining what an agent can produce structurally, rather than requiring a human to review every individual output. This reframes governance from an inspection problem (a person checks every agent action) to a constraint problem (the structure of the work limits what the agent can do in the first place) — a shift with direct operational leverage, since inspection does not scale with agent throughput but structural constraint does.

7.2 Natural-language instruction is not a control mechanism

The strongest single empirical argument for this section's thesis is negative: instructing agents in natural language to follow good practice does not reliably work, which is precisely why structural constraints are necessary rather than optional. Ouatiti et al.'s study of 4,550 agentic pull requests across 81 open-source repositories found that “explicit logging instructions are rare (4.7%) and ineffective, as agents fail to comply with constructive requests 67% of the time,” with humans performing “72.5% of post-generation log repairs, acting as 'silent janitors' who fix logging and observability issues without explicit review feedback” [6.5]. The authors' own conclusion states the governance implication directly: this “dual failure in natural language instruction... suggest[s] that deterministic guardrails might be necessary to ensure consistent logging practices” [6.5]. This finding generalizes well beyond logging: if agents do not reliably follow explicit natural-language instructions even when those instructions are given, then a governance policy document — however well written — cannot be the control mechanism. The control has to be structural: a test the code must pass, a bounded context the code cannot cross, an architecture decision record the agent must conform to.

7.3 Structural constraints as measurable governance controls

Recent 2026 software-engineering research supplies direct quantitative evidence that structural, engineering-native constraints outperform natural-language governance and manual review, and does so specifically in the context of AI coding agents. Lipsanen et al.'s Shift-Up framework reinterprets “established software engineering practices, like executable requirements (BDD), architectural modeling (C4), and architecture decision records (ADRs), as structural guardrails for GenAI-native development,” finding that “embedding machine-readable requirements and architectural artifacts stabilizes agent behavior, reduces implementation drift, and shifts human effort toward higher-level design and validation activities” [6.2]. Their framing of the alternative failure mode is precise: unstructured “vibe coding” — prompting an agent without embedded structural constraints — “often suffers from architectural drift, limited traceability, and reduced maintainability” [6.2], exactly the failure mode governance-by-inspection alone cannot reliably catch at agent speed and volume.

Madiraju and Madiraju's RigorBench supplies the clearest quantitative payoff claim available: structured process discipline “not only improves process quality scores by an average of 41% but also raises downstream outcome correctness by 17%, providing the first quantitative evidence that how agents code matters as much as what they produce” [6.3]. Their critique of outcome-only evaluation is directly relevant to governance design: “an agent that arrives at a correct solution through reckless trial-and-error, without planning, verification, or graceful recovery, is fundamentally less reliable than one that follows sound engineering discipline” [6.3] — a correct output today is not evidence of a governable process tomorrow, at a different scale or under different conditions.

7.4 Domain-Driven Design: structure that constrains, but does not automate, meaning

Domain-Driven Design's contribution — ubiquitous language, bounded contexts, and invariants — is directly continuous with the ontology-resolution argument of Section 4, and the recent literature specifically warns against treating DDD as automatable, which strengthens rather than weakens its governance value. Eisenreich, Jusic, and Wagner's evaluation of an LLM-based DDD prompting framework against real enterprise requirements found that early steps — establishing ubiquitous language, simulating event storming, identifying bounded contexts — "worked well," but that "accumulated errors rendered the artifacts generated from Steps 4 and 5 [aggregate design and technical architecture mapping] impractical" [6.8]. Their conclusion is worth quoting for its precision: "LLMs can enhance, but not replace, architectural expertise, offering a practical tool to reduce the effort and overhead of DDD while preserving human-centric decision-making" [6.8]. This is a negative result in the narrow sense (full automation of DDD failed) but a positive one for this article's argument in the broader sense: DDD's governance value comes specifically from being a human-authored structural constraint that an agent then operates within, not a specification an agent can generate wholesale for itself — the same logic, applied to code architecture, as the ontology audit's insistence on a named accountable owner for each contested definition (Section 4).

Where DDD constraints are formalized precisely enough, they can be enforced automatically once established: Dinçoğuz, Kendir, and Karakaya's DDD-Enforcer, a real-time analysis tool combining abstract-syntax-tree inspection with LLM-based checking, achieved "100% detection accuracy across 15 violation cases with an average latency of 4.49 seconds" against ubiquitous-language and bounded-context violations [6.9] — demonstrating that once a human has done the DDD design work, enforcement of that design against agent-authored code can itself be automated and run continuously, closing the loop between human-authored structure and machine-scale governance.

7.5 Test-first discipline as executable specification

Extreme Programming's test-first discipline supplies the third leg of this section's argument: a test written before code is generated functions simultaneously as a specification and as a verification mechanism for agent output. Mathews and Nagappan's foundational study of test-driven development applied to LLM-based code generation found that “including test cases leads to higher success in solving programming challenges,” concluding that “TDD is a promising paradigm for helping ensure that the code generated by LLMs effectively captures the requirements” [6.7]. A complementary 2025 benchmark treating tests as the primary prompt input, evaluated across nineteen frontier models, found that “instruction following and in-context learning” — not general coding proficiency or pretraining knowledge — were the critical capabilities for TDD success [6.7] — a finding consistent with this section's broader claim that structural specification, not model capability alone, is what determines whether agent-authored code stays governable.

7.6 Refactoring evidence: agents do local work, humans supply design intent

A final piece of evidence supports the specific division of labor this section proposes — engineering structure supplied by humans, execution constrained by that structure performed by agents. Horikawa et al.'s empirical study of 15,451 refactoring instances across 12,256 agent-authored pull requests found that agentic refactoring is common (targeted explicitly in 26.1% of commits) but “dominated by low-level, consistency-oriented edits... reflecting a preference for localized improvements over the high-level design changes common in human refactoring” [6.6]. This is exactly the division of labor Section 2's partition move argues for, restated in engineering terms: agents reliably execute structured, explicit, bounded work (renaming variables, adjusting parameter types) but do not reliably originate architectural intent, which must continue to come from a human — reinforcing DDD's role as a human-supplied structural constraint rather than an agent-generated one.

7.7 A technical-layer echo of the same logic

Recent research on ontology-grounded agent architectures — Tuan and Sanyal's three-layer Role, Domain, and Interaction ontology, introduced in Section 4 — makes the same partition-and-guardrail logic explicit at the model layer, proposing structural constraints specifically “to constrain LLM reasoning inside enterprise agents” [3.1, 6.1] and addressing “hallucination, domain drift, and the inability to enforce regulatory compliance at the reasoning level” [6.1] through architecture rather than review. Shamsujjoha et al.'s Swiss Cheese Model for AI safety extends James Reason's defence-in-depth concept — the architectural sibling of Perrow's coupling analysis from Section 5 — into a reference architecture for “multi-layered guardrails” spanning “quality attributes, pipelines, and artifacts” [6.10], on the grounds that “designing effective runtime guardrails is challenging due to the agents' autonomous and non-deterministic behavior” [6.10]. Taken together, Sections 4 and 7 describe the same underlying discipline — resolve ambiguity structurally, once, at the level of definitions and architecture, rather than repeatedly, at the level of individual review — applied first to organizational semantics and then to code.

8. Synthesis: The AO Governance Map

Put together, the six moves form a single translation map: existing AO Method vocabulary, developed for human decision rights and organizational design, applied directly to AI agent governance rather than requiring the organization to learn a parallel framework from scratch.

AO Element	AI Governance Translation	Primary literature support
Decision-flow diagnostic	Map which decisions are agent-eligible versus human-owned, based on ambiguity and accountability rather than task type.	Sun [1]; Lubars & Tan [1.2]; EU AI Act Art. 14 [1.4]; NIST AI RMF App. C [1.6]
Autonomy-risk boundary	Reversibility, blast radius, and correlation gate the level of agent autonomy; static risk tiers are a floor, not a ceiling.	Zheng et al. [2.1]; Feng et al. [2.2]; Engin & Hand [2.3]; Mitchell et al. [2.4]
Ontology diagnostic	Resolve entity definitions across systems, with a named accountable owner, before agents inherit and amplify them.	Tuan & Sanyal [3.1]; Semarchy [3.7]; IBM [3.8]; Qiang et al. [3.6]
Coupling pace	Explicit social brakes — sampling review, exception escalation — between agent execution speed and human review cadence.	Pidgeon/Perrow [4.1]; Macrae [4.2]; Renieris et al. [4.5]; JAPCC [4.6]
Hoshin spine	Annual guardrails plus monthly rolling catchball review; continuous validation rather than a static policy document.	Tennant & Roberts [5.1]; Ćirić et al. [5.2]; Lee et al. [5.4]; Bakirtzis et al. [5.5]
Software-method guardrails	XP, FDD, and DDD constrain agent-authored execution structurally; enforcement can itself be automated once design intent is human-authored.	Lipsanen et al. [6.2]; Madiraju & Madiraju [6.3]; Ouatiti et al. [6.5]; Eisenreich et al. [6.8]

The map's underlying claim is not that any single literature independently validates the AO Method. It is that six independently developed bodies of research — regulatory scholarship on human oversight, agentic-AI risk-tiering research, enterprise ontology and knowledge-graph engineering, sociotechnical accident theory, strategy-deployment methodology, and software-engineering discipline — converge on the same underlying diagnosis from six different starting points: that governing AI agents effectively requires organizational-design decisions the compliance-extension paradigm identifies as necessary but does not itself specify. The AO Method's contribution is not a new theory of any one of these six domains; it is a single, already-tested organizational vocabulary that lets a firm apply its existing decision-rights, autonomy, and strategy-deployment discipline to all six at once, rather than adopting six separate specialist frameworks that do not share a common language.

9. Discussion and Limitations

9.1 What this article has, and has not, established

This article has argued, and supported with source-cited literature across six domains, that each of the AO Method's governance moves has an independently developed analogue in current academic and industry research, and that the overall thesis — AI governance is an organizational-design problem, not primarily a compliance-extension problem — is echoed across regulatory scholarship [1.5, G3], organizational theory [1, G1], safety science [4.2, 4.3], and software engineering [6.2, 6.5]. This is evidence of convergent plausibility: independent research communities, working from different starting assumptions and different literatures, have reached compatible conclusions about where AI governance's real difficulty lies. It is not evidence that the AO Method's specific implementation of these six moves — as opposed to the general diagnosis they share — produces better governance outcomes than alternative organizational-design approaches, or than a well-executed version of the compliance-extension paradigm itself.

Several limitations follow directly from this distinction.

9.2 The AO Method's own evidence base is not yet independently validated

The AO Method itself, as a decision-rights and autonomy-calibration framework, predates this article's AI governance application and was developed through consulting practice rather than controlled research. This article has grounded each of its six moves in external literature, but that literature validates the component ideas (partition by ambiguity, autonomy calibration by reversibility and blast radius, ontology resolution, coupling-pace design, adaptive strategy deployment, engineering-constraint governance) rather than the AO Method's specific combination and sequencing of them. Sun's decision-rights framework [1], for instance, is validated through “two worked applications” [1] — a modest evidence base by the standards of, say, Ouatiti et al.'s 4,550-pull-request empirical study [6.5]. Readers should treat the individual literature citations in Sections 2 through 7 as considerably stronger evidence than the overall synthesis in Section 8, which remains, at this stage, an argument for structural coherence rather than a demonstrated causal claim.

9.3 Hoshin Kanri's evidentiary limits transfer to Move Five

Section 6 already flagged this directly: Wilson, Cudney, and Marley's systematic review found “relatively little published research” on Hoshin Kanri despite decades of Western use [5.3], and Ćirić et al. found that available case studies mostly lack “concrete quantitative results” and sufficient longitudinal data to establish long-term effectiveness [5.2]. Because Move Five borrows Hoshin Kanri's cadence and catchball mechanism directly, this evidentiary gap in the source methodology transfers to the AO Method's AI governance application. The stronger, more recent evidence for continuous governance specifically — Lee et al.'s Co-Lifecycle Governance framework [5.4], Bakirtzis et al.'s dynamic certification proposal [5.5] — comes from adjacent literatures (medical AI regulation, AI certification theory) that independently support the principle of continuous, threshold-triggered review, but neither directly tests the Hoshin Kanri mechanism itself in an AI governance context. Empirical work specifically evaluating rolling catchball cadences against AI-agent governance outcomes, as distinct from evaluating either Hoshin Kanri in its traditional manufacturing context or continuous governance frameworks that do not use catchball, remains an open gap.

9.4 Access and coverage gaps in the literature review itself

Two sources initially identified for this review — Williams and Yampolskiy's analysis of AI failure modes and Ibitoye, Nkwo, and Orji's work in AI and Ethics — could not be verified through direct publisher access during this research and were excluded rather than cited from memory, consistent with this article's evidence discipline. Two IEEE conference papers on automating and enforcing Domain-Driven Design (Eisenreich et al. [6.8]; Dinçoğuz et al. [6.9]) were accessible only through academic index records rather than direct publisher fetch; their metadata and quoted abstracts are reported with that caveat. These are minor gaps against a 42-source base, but they illustrate a more general limitation: rapidly moving arXiv preprints dominate several sections of this review (particularly Sections 3, 5, and 7, covering 2025-2026 agentic-AI research), and preprints have not undergone peer review at the time of citation. Where this article cites a specific quantitative claim from a preprint — Tuan and Sanyal's $p < .001$ result [3.1], Madiraju and Madiraju's 41%/17% figures [6.3], Ouatiti et al.'s 4.7%/67% figures [6.5] — readers should treat these as promising but not yet peer-reviewed evidence, pending eventual publication and independent replication.

9.5 The compliance-extension paradigm is not refuted, only insufficient alone

This article has argued that the organizational-design paradigm addresses a gap the compliance-extension paradigm leaves open, not that the compliance-extension paradigm is wrong or dispensable. Coglianese and Crum's management-based regulation [G3], the EU AI Act's Article 14 [1.4], and NIST's AI RMF [1.6] remain the mechanisms by which external accountability is created and by which organizations are compelled to build internal governance capacity in the first place. Engin and Hand's explicit position — that dimensional governance "complements rather than replaces" categorical frameworks [2.3] — is the correct relationship to draw between the two paradigms generally: statutory risk tiers and human-oversight mandates set a floor and create external pressure; organizational-design methods like the AO Method's six moves determine whether an organization actually meets that floor in a way that holds up under real operating conditions, or merely produces the "surface-level" responsible-AI practice the BCG survey data suggests is currently the norm for a majority of organizations [G5].

9.6 Directions for further validation

Three lines of empirical work would materially strengthen the case this article makes. First, longitudinal case studies of organizations applying decision-rights partitioning, autonomy calibration, and ontology audits specifically to agentic AI deployments, tracking incident rates, escalation frequency, and time-to-detection of governance drift against comparable organizations using compliance-extension governance alone. Second, controlled comparison of Hoshin-style rolling catchball cadences against both static annual review and Lee et al.'s Co-Lifecycle Governance pillars [5.4], to establish whether the specific catchball mechanism adds value over threshold-triggered surveillance alone. Third, replication of the software-engineering findings cited in Section 7 — particularly RigorBench's process-discipline effect sizes [6.3] and the DDD automation-limit finding [6.8] — across a wider range of agent architectures and enterprise domains than the current, still-narrow 2025-2026 evidence base covers. Until this work is done, this article's contribution should be read as a well-evidenced argument for where to look for AI governance solutions — organizational design rather than compliance extension — and a coherent, literature-grounded translation of one organization-design method into that space, rather than as a validated intervention with demonstrated outcomes of its own.

10. Conclusion

AI governance built by importing IT risk and compliance frameworks wholesale rarely survives sustained contact with a real organization, and the adoption data available in 2025-2026 show why: broad implementation, shallow maturity, and internal actors who already recognize the accountability is theirs to hold, but frequently lack the organizational-design vocabulary to hold it well. This article has argued that the Adaptive Organization Method's existing tools for decision rights, autonomy calibration, and strategic cadence transfer to AI governance with limited modification, because the underlying problem does not change when one of the decision-making agents in an organization is a machine: someone still has to own ambiguous judgment, someone still has to calibrate how much unsupervised action is acceptable given what could go wrong, someone still has to resolve what “customer” or “done” actually means before an automated system amplifies whatever ambiguity is already there, and someone still has to design how fast systems and slow oversight connect to each other. Six independently developed literatures — spanning regulatory scholarship, agentic-AI risk research, enterprise ontology engineering, sociotechnical safety science, strategy-deployment methodology, and software engineering — converge on this same underlying diagnosis from six different directions. The AO Method's contribution is to give organizations a single, already-practiced vocabulary for acting on that diagnosis: governance as design, diagnosed and owned by the people who already run the system, rather than governance as a document written once and quietly set aside the moment real work collides with it.

References

[1] Sun, F. (2026). Authority and Responsibility Allocation in Human–AI Collaborative Decision-Making: Governance Mechanisms for Enterprises. ICCK Transactions on Systems Safety and Reliability, 2(3), 192–206. https://doi.org/10.62762/TSSR.2026.930520

[1.2] Lubars, B., & Tan, C. (2019). Ask Not What AI Can Do, But What AI Should Do: Towards a Framework of Task Delegability. NeurIPS 2019. https://arxiv.org/abs/1902.03245

[1.3] Hemmer, P., Westphal, M., Schemmer, M., Vetter, S., Vössing, M., & Satzger, G. (2023). Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. 28th International Conference on Intelligent User Interfaces (IUI '23). https://arxiv.org/abs/2303.09224

[1.4] European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 14 — Human Oversight. https://artificialintelligenceact.eu/article/14/

[1.5] Enqvist, L. (2023). "Human oversight" in the EU artificial intelligence act: what, when and by whom? Law, Innovation and Technology, 15(2), 508–535. https://doi.org/10.1080/17579961.2023.2245683

[1.6] National Institute of Standards and Technology. (2023). AI Risk Management Framework (AI RMF 1.0), Appendix C: AI Risk Management and Human-AI Interaction. U.S. Department of Commerce. https://airc.nist.gov/airmf-resources/airmf/appendices/app-c-ai-risk-management-and-human-ai-interaction/

[2.1] Zheng, H., Dong, Q., Depena, R. K., Bhatia, J. D., Xiao, F., & Xu, P. (2026). Separating Capability from Permission: A Governance Framework for Agentic AI Autonomy Levels. https://arxiv.org/abs/2607.23438

[2.2] Feng, K. J. K., McDonald, D. W., & Zhang, A. X. (2025). Levels of Autonomy for AI Agents. https://arxiv.org/abs/2506.12469

[2.3] Engin, Z., & Hand, D. (2026). Towards adaptive categories: Dimensional governance for agentic AI. Data & Policy, 8. Cambridge University Press. https://www.cambridge.org/core/product/identifier/S2632324926100832/type/journal_article

[2.4] Mitchell, M., Ghosh, A., Luccioni, A. S., & Pistilli, G. (2025). Fully Autonomous AI Agents Should Not be Developed. https://arxiv.org/abs/2502.02649

[2.5] Cihon, P., Stein, M., Bansal, G., Manning, S., & Xu, K. (2025). Measuring AI agent autonomy: Towards a scalable approach with code inspection. https://arxiv.org/abs/2502.15212

[2.6] Raza, S., Sapkota, R., Karkee, M., & Emmanouilidis, C. (2025). TRiSM for Agentic AI: A Review of Trust, Risk, and Security Management in LLM-based Agentic Multi-Agent Systems. https://arxiv.org/abs/2506.04133

[2.7] European Union. (2024). Regulation (EU) 2024/1689 (Artificial Intelligence Act), Article 6 — Classification Rules for High-Risk AI Systems. https://artificialintelligenceact.eu/article/6/

[2.8] Gartner. (2025, June 25). Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027 [Press release]. https://www.gartner.com/en/newsroom/press-releases/2025-06-25-gartner-predicts-over-40-percent-of-agentic-ai-projects-will-be-canceled-by-end-of-2027

[2.9] Singla, A., Sukharevsky, A., Hall, B., Yee, L., Chui, M., & Balakrishnan, T. (2025). The state of AI in 2025: Agents, innovation, and transformation. QuantumBlack, AI by McKinsey / McKinsey & Company. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

[3.1] Tuan, T. L., & Sanyal, A. (2026). Ontology-Constrained Neural Reasoning in Enterprise Agentic Systems: A Neurosymbolic Architecture for Domain-Grounded AI Agents. https://arxiv.org/abs/2604.00555

[3.2] Laurenzi, E., Mathys, A., & Martin, A. (2024). An LLM-Aided Enterprise Knowledge Graph (EKG) Engineering Process. Proceedings of the AAAI Symposium Series, 3(1), 148–156. https://doi.org/10.1609/aaaiss.v3i1.31194

[3.3] Shimizu, C., & Hitzler, P. (2024). Accelerating Knowledge Graph and Ontology Engineering with Large Language Models. https://arxiv.org/abs/2411.09601

[3.4] Feng, X., Wu, X., & Meng, H. (2024). Ontology-grounded Automatic Knowledge Graph Construction by LLM under Wikidata schema. HI-AI@KDD 2024; CEUR Workshop Proceedings 3841, 117–135. https://arxiv.org/abs/2412.20942

[3.5] Rasmussen, P., Paliychuk, P., Beauvais, T., Ryan, J., & Chalef, D. (2025). Zep: A Temporal Knowledge Graph Architecture for Agent Memory. https://arxiv.org/abs/2501.13956

[3.6] Qiang, Z., Wang, W., & Taylor, K. (2023–2026). Agent-OM: Leveraging LLM Agents for Ontology Matching. VLDB 2025 / OM 2025. https://arxiv.org/abs/2312.00326

[3.7] Semarchy. (2025, April 11). 74% of Businesses Will Invest in AI this Year – But Data Quality Concerns Remain [Press release]. https://semarchy.com/press-releases/ai-data-quality-gap-study/

[3.8] Krantz, T., & Jonker, A. (2026). The True Cost of Poor Data Quality. IBM Think Insights. https://www.ibm.com/think/insights/cost-of-poor-data-quality

[4.1] Pidgeon, N. (2011). In retrospect: Normal Accidents. Nature, 477, 404–405. https://doi.org/10.1038/477404a

[4.2] Macrae, C. (2021/2022). Learning from the Failure of Autonomous and Intelligent Systems: Accidents, Safety, and Sociotechnical Sources of Risk. Risk Analysis. https://doi.org/10.1111/risa.13850

[4.3] Dobbe, R. (2025). AI Safety is Stuck in Technical Terms — A System Safety Response to the International AI Safety Report. https://arxiv.org/abs/2503.04743

[4.4] Kilian, K. A. (2024/2025). Beyond Accidents and Misuse: Decoding the Structural Risk Dynamics of Artificial Intelligence. https://arxiv.org/abs/2406.14873

[4.5] Renieris, E. M., Kiron, D., Mills, S., & Kleppe, A. (2025, September 16). Agentic AI at Scale: Redefining Management for a Superhuman Workforce. MIT Sloan Management Review / Boston Consulting Group. https://sloanreview.mit.edu/article/agentic-ai-at-scale-redefining-management-for-a-superhuman-workforce/

[4.6] Joint Air Power Competence Centre. (2021). Is Human-On-the-Loop the Best Answer for Rapid Relevant Responses? https://www.japcc.org/essays/is-human-on-the-loop-the-best-answer-for-rapid-relevant-responses/

[5.1] Tennant, C., & Roberts, P. (2001). Hoshin Kanri: a tool for strategic policy deployment. Knowledge and Process Management, 8(4), 262–269. https://doi.org/10.1002/kpm.121

[5.2] Ćirić, M., Tomašević, I., Stojanović, D., & Slović, D. Systematic Literature Review on Hoshin Kanri Methodology and Application in the Service Sector. University of Belgrade Faculty of Organizational Sciences. https://rfos.fon.bg.ac.rs/bitstream/123456789/3000/2/Ciric-Tomasevic-Stojanovic-Slovic.pdf

[5.3] Wilson, R., Cudney, E. A., & Marley, R. J. (2023). Current Status of Hoshin Kanri. The TQM Journal. https://doi.org/10.1108/TQM-07-2022-0216

[5.4] Lee, J. H., Choi, B., Jeong, K., Suh, S. W., Kim, J. H., & Son, D.-S. (2026). Co-Lifecycle Governance for Learning Medical AI. Journal of Medical Internet Research, 28, e90654. https://doi.org/10.2196/90654

[5.5] Bakirtzis, G., Aler Tubella, A., Theodorou, A., Danks, D., & Topcu, U. (2024). Navigating the sociotechnical labyrinth: Dynamic certification for responsible embodied AI. https://arxiv.org/abs/2409.00015

[6.1] Tuan, T. L., & Sanyal, A. (2026). Ontology-Constrained Neural Reasoning in Enterprise Agentic Systems. (See also [3.1].) https://arxiv.org/abs/2604.00555

[6.2] Lipsanen, P., Rannikko, L., Christophe, F., Kalliokoski, K., Stirbu, V., & Mikkonen, T. (2026). Shift-Up: A Framework for Software Engineering Guardrails in AI-native Software Development. https://arxiv.org/abs/2604.20436

[6.3] Madiraju, M. B., & Madiraju, M. S. P. (2026). RigorBench: Benchmarking Engineering Process Discipline in Autonomous AI Coding Agents. https://arxiv.org/abs/2606.22678

[6.4] Navneet, S. K., & Chandra, J. (2025). Rethinking Autonomy: Preventing Failures in AI-Driven Software Engineering. https://arxiv.org/abs/2508.11824

[6.5] Ouatiti, Y. E., Sayagh, M., Li, H., & Hassan, A. E. (2026). Do AI Coding Agents Log Like Humans? An Empirical Study. https://arxiv.org/abs/2604.09409

[6.6] Horikawa, K., Li, H., Kashiwa, Y., Adams, B., Iida, H., & Hassan, A. E. (2025). Agentic Refactoring: An Empirical Study of AI Coding Agents. https://arxiv.org/abs/2511.04824

[6.7] Mathews, N. S., & Nagappan, M. (2024). Test-Driven Development for Code Generation. ASE 2024. Complementary source: Cui, Y. (2025). Tests as Prompt: A Test-Driven-Development Benchmark for LLM Code Generation. https://arxiv.org/abs/2402.13521

[6.8] Eisenreich, T., Jusic, H., & Wagner, S. (2026). Automating Domain-Driven Design: Experience with a Prompting Framework. ICSA-C 2026. https://doi.org/10.1109/ICSA-C68850.2026.00085

[6.9] Dinçoğuz, A., Kendir, A., & Karakaya, M. (2026). DDD-Enforcer: An AI-Powered Multi-Agent System for Real-Time Domain-Driven Design Enforcement. IISEC 2026. https://doi.org/10.1109/IISEC69317.2026.11418529

[6.10] Shamsujjoha, M., Lu, Q., Zhao, D., & Zhu, L. (2024/2025). Swiss Cheese Model for AI Safety: A Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model Based Agents. https://doi.org/10.48550/arXiv.2408.02205

[G1] Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022/2023). Putting AI Ethics into Practice: The Hourglass Model of Organizational AI Governance. https://arxiv.org/abs/2206.00335

[G2] Schneider, J., Abraham, R., Meske, C., & vom Brocke, J. (2022). Artificial Intelligence Governance for Businesses. Information Systems Management. https://arxiv.org/abs/2011.10672

[G3] Coglianese, C., & Crum, C. R. (2024). Taking Training Seriously: Human Guidance and Management-Based Regulation of Artificial Intelligence. https://arxiv.org/abs/2402.08466

[G4] Reuel, A., Bucknall, B., Casper, S., et al. (2024/2025). Open Problems in Technical AI Governance. Transactions on Machine Learning Research, 2025. https://arxiv.org/abs/2407.14981

[G5] Rulf, K., Sieben, M., Mills, S., & Kleppe, A. (2026, April 30). Responsible AI Needs More Than Good Intentions. Boston Consulting Group. https://www.bcg.com/publications/2026/responsible-ai-needs-more-than-good-intentions