# The Competitive Deployment Dilemma: Market Incentives, Collective Action, and the Policy Conditions for Accountable AI Authority

*Paper 4 in a Series on AI Governance and Organizational Authority*

---

Alexander Huseby
Founder, Cognitive Liberty Institute
Oslo, Norway — alexander@coglib.no
2026

---

Abstract

*The three preceding papers in this series established that AI decision-making authority is expanding faster than accountability frameworks can keep pace, applied this finding to documented enterprise deployments, and specified the technical architecture required to maintain meaningful human oversight at each level of the AI Authority Maturity Model (AAMM). This paper addresses the structural question that underlies the governance gap: why does it persist, and under what conditions will it close? The argument is that the governance gap is not primarily a failure of organizational ethics or awareness. It is a collective action problem in the precise sense articulated by Olson (1965): individual organizations rationally underinvest in AI governance because the costs are private and immediate, while the benefits — avoided systemic risk, maintained institutional trust, public confidence in AI systems — are shared and diffuse. No individual organization has sufficient incentive to bear the full cost of accountability architecture when competitors can deploy without it and free-ride on the resulting institutional trust. This structural diagnosis implies that governance investment will remain systematically below socially optimal levels unless structural conditions change. This paper examines three mechanisms that can change those conditions: mandatory minimum standards enforced through regulation, market mechanisms that price governance quality differentially, and polycentric coordination that creates shared governance infrastructure. Drawing on the EU AI Act's phased enforcement timeline, the emergence of AI liability insurance and ISO/IEC 42001 certification markets, and the limitations of voluntary commitments demonstrated by the Seoul Frontier AI Safety Commitments (2024), the paper concludes by specifying the conditions under which governance investment becomes self-sustaining as a source of competitive advantage rather than a competitive burden.*

Keywords: collective action, AI governance, competitive dynamics, EU AI Act, AI liability insurance, ISO 42001, voluntary commitments, regulatory policy, polycentric governance, governance premium, AAMM, free-rider problem

---

## 1. Introduction

The three preceding papers in this series have built a layered argument. Paper 1 (Huseby, 2025a) introduced the AI Authority Maturity Model (AAMM) and established that the trajectory toward expanded AI decision-making authority is economically rational, incrementally normalized, and systematically ahead of the accountability architecture required to make it responsible. Paper 2 (Huseby, 2026a) applied the AAMM to four documented enterprise deployments — BlackRock, JPMorgan Chase, Maersk, and Unilever — and found the governance-deployment gap to be consistent across industries. Paper 3 (Huseby, 2026b) descended to the engineering level, specifying the technical mechanisms — domain bounding, HITL interrupt architecture, circuit breakers, tamper-evident audit trails — required to maintain meaningful human oversight at each AAMM level.

Those three papers addressed what the governance gap is, where it manifests, and what technical architecture is required to close it. This paper addresses the prior question: why does the gap persist despite widespread awareness of it, and under what structural conditions will it close?

The answer proposed here is that the governance gap is a collective action problem. Individual organizations do not underinvest in AI governance because they are unaware of its importance, nor because they are indifferent to systemic risk. They underinvest because the incentive structure they face makes underinvestment rational. The costs of governance — engineering overhead, deployment latency, ongoing audit costs, organizational complexity — are borne by the individual organization immediately. The benefits — avoided systemic failures, maintained public trust in AI systems, reduced tail risk across the institutional landscape — are shared across all organizations and accrue slowly. Under these conditions, the individually rational choice is to deploy AI with minimal governance and free-ride on the accountability infrastructure that more cautious competitors bear the cost of building.

This diagnosis has direct implications for the policy prescriptions that follow. If the governance gap were a knowledge problem, better guidelines and frameworks would close it. If it were an ethics problem, better values and intentions would close it. Because it is a structural incentive problem, closing it requires structural changes to the incentive environment: regulation that makes governance mandatory and levels the competitive playing field, market mechanisms that make governance quality visible and differentially rewarded, and coordination that creates shared governance infrastructure so individual organizations do not bear the full cost alone.

## 2. The Structural Problem: Competitive Deployment as Collective Action Failure

Mancur Olson's foundational work, *The Logic of Collective Action* (1965), established that large groups of individuals sharing a common interest will not spontaneously act to advance that interest, even when doing so would benefit all members. The reason is the free-rider problem: the benefits of collective action,

once produced, are non-excludable — they accrue to all members regardless of whether they contributed to their production. A rational individual therefore has an incentive to let others bear the cost of collective action and benefit from the result without contributing. When all individuals reason this way, the collective good is underprovided.

The AI governance gap fits this structure precisely. The collective good in question is institutional trust in AI systems: the shared confidence of regulators, customers, employees, and the public that AI-assisted organizational decisions are accountable, auditable, and subject to meaningful human oversight. This trust, once established, is non-excludable — an organization that deploys AI recklessly benefits from the institutional trust created by competitors who invest in governance, without contributing to that trust. The rational response to this structure, for any individual organization facing competitive pressure, is to minimize governance investment and free-ride on the accountability credibility that more cautious organizations establish.

This is not a theoretical abstraction. The McKinsey (2025) finding that fewer than 25% of companies have board-approved, structured AI policies, despite 88% using AI in core business functions, is the empirical signature of the free-rider dynamic: the majority of organizations are deploying AI without accountability infrastructure while benefiting from the regulatory and reputational environment established by the minority that invests in governance. The MIT CISR finding (Weill, Woerner & Banner, 2025) that AI-savvy boards outperform by 10.9 percentage points in ROE suggests the long-run governance premium is real — but the short-run competitive pressure operates on a different timescale than the long-run benefit, and the rational response to short-run competitive pressure is to deploy fast and govern later.

Olson identified two solutions to collective action problems in large groups: coercion — making contribution mandatory so that free-riding is not possible — and selective incentives — rewards or penalties that apply specifically to participants and non-participants respectively, not to the group as a whole. Both mechanisms appear in the emerging AI governance landscape, and both are necessary. The sections that follow examine each in turn.

A further dimension of the collective action problem operates at the inter-organizational level: the competitive dynamics between organizations in different regulatory jurisdictions create a potential race to the bottom in governance standards. An organization subject to stringent EU AI Act requirements competes with organizations operating under less demanding regulatory regimes. If governance is a cost and not a differentiator, the competitive pressure favors the less regulated competitor. This inter-jurisdictional dimension of the collective action problem requires international coordination mechanisms that operate at a different level from the organizational governance frameworks examined in the preceding papers.

## 3. Why Voluntarism Is Necessary but Insufficient

The dominant governance response to the AI governance gap at the international level has been voluntary commitment. The Bletchley Declaration (November 2023), in which 28 governments committed to deepening cooperation on AI risks, and the Frontier AI Safety Commitments (Seoul Summit, May 2024), in which 16 leading AI companies committed to responsible development practices including halting models that present extreme unmitigated risks, represent the most prominent examples of this approach.

These commitments are not without value. They establish norms, create reputational expectations, and build the social infrastructure for more binding governance over time. The Seoul commitments in particular were notable for their geographic breadth — including signatories from North America, Europe, Asia, and the Middle East — and for their specific commitment to publish threshold definitions for intolerable AI risks. Yoshua Bengio, one of the leading figures in AI safety research, described them as "an important step forward in establishing an international governance regime to promote AI safety" while noting they would "have to be accompanied by other regulatory measures" (Computer Weekly, 2024).

The structural limitation of voluntarism is precisely what Olson's analysis predicts. Voluntary commitments create no mechanism for excluding free-riders from the benefits of collective governance. An AI company that does not sign the Seoul commitments can still benefit from the reputational infrastructure that signatories create: if the 16 signing companies establish a norm of responsible AI development, the reputational landscape within which all AI companies operate improves, regardless of whether they contributed to establishing that norm. The selective incentive — a specific reward for signatories that non-signatories cannot access — is absent.

The voluntarism gap is also empirically visible. The Seoul commitments covered 16 frontier AI lab companies. The universe of organizations deploying AI in consequential organizational contexts — the enterprise AI deployers examined in Paper 2 — is orders of magnitude larger and almost entirely outside any voluntary commitment framework. The frontier lab is a small, well-organized group for whom collective action is relatively tractable, as Olson predicts. Enterprise AI deployers are a large, heterogeneous group with diverse interests, limited coordination mechanisms, and no natural organizing principle for collective governance action — exactly the type of group for which Olson's analysis predicts collective action will systematically fail without external structural intervention.

This is not an argument against voluntary commitments. It is an argument that voluntary commitments are a necessary but not sufficient component of the governance ecosystem, and that the institutional infrastructure they create is most valuable as a foundation for subsequent mandatory mechanisms rather than as a substitute for them.

## 4. Mechanism 1: Mandatory Standards and the Competitive Floor

The most direct structural solution to the collective action problem is coercion in Olson's sense: mandatory governance requirements that remove the free-rider advantage of underinvestment by making governance a condition of market access rather than a voluntary expenditure. Mandatory standards level the competitive playing field by ensuring that the cost of governance is borne by all market participants rather than only by those organizations willing to accept a competitive disadvantage in exchange for institutional accountability.

The EU AI Act (Regulation 2024/1689) is the most consequential mandatory standard currently in force. Its phased implementation schedule is now in active enforcement: prohibited AI practices became enforceable on February 2, 2025, with penalties of up to €35 million or 7% of global annual revenue. General-Purpose AI model obligations became mandatory on August 2, 2025. High-risk AI system requirements — covering AI used in employment decisions, credit scoring, education, and law enforcement, among other contexts — are enforceable from August 2, 2026, with providers required to complete conformity assessments, register systems in the EU AI database, implement quality management systems, and maintain human oversight measures as specified in Article 14 (Secure Privacy, 2026).

The Act's significance as a governance mechanism extends beyond its direct regulatory requirements. The GDPR precedent is instructive: enacted in 2018 with application to any organization processing EU personal data regardless of where it is headquartered, GDPR effectively set a global data protection standard because multinational organizations found it more efficient to apply GDPR-level protections globally than to maintain jurisdiction-specific practices. The EU AI Act, with its extraterritorial reach to any organization placing AI systems on the EU market or using AI outputs within the EU, has the same structural potential to establish a global accountability floor. An organization that complies with the EU AI Act's high-risk requirements has, by construction, implemented governance architecture that meets or exceeds what most jurisdictions currently require.

The mandatory standard mechanism addresses the collective action problem by converting governance from a voluntary expenditure that creates competitive disadvantage into a mandatory cost that all market participants share. When governance is mandatory, the free-rider advantage of non-compliance is eliminated: competitors who deploy AI without accountability architecture are not competing on equal terms but operating outside the legal conditions of market access. The competitive playing field is leveled at the governance floor, and competition then occurs on capability, performance, and efficiency within that floor.

The limitation of the mandatory standard mechanism is its geographic scope and enforcement capacity. Organizations outside the EU's jurisdictional reach, or those operating in sectors not classified as high-risk under the Act's risk taxonomy, face no mandatory accountability requirements. The inter-jurisdictional race to the bottom dynamic identified in Section 2 is partially mitigated by the GDPR-precedent effect but not eliminated. Organizations in jurisdictions without mandatory AI governance requirements can compete with

EU-compliant organizations on governance cost, creating a persistent structural disadvantage for compliant organizations in competitive markets where jurisdiction is ambiguous.

## 5. Mechanism 2: Market Mechanisms and Governance Pricing

Olson's second solution to collective action problems is selective incentives: rewards or penalties that apply specifically to participants in governance, creating a private incentive for compliance independent of the collective benefit. In the AI governance context, two market mechanisms are emerging that function as selective incentives: AI liability insurance markets that price governance quality differentially, and AI management system certification schemes that make governance quality verifiable and marketable.

### 5.1 AI Liability Insurance

The AI liability insurance market is nascent but growing. Relm Insurance, a Bermuda-domiciled specialty insurer, announced three AI-specific policies in January 2025: NOVAAI (cyber and tech errors and omissions for AI platform companies), PONTAAI (excess difference in conditions policy for organizations with third-party liability from AI use or development), and a third policy addressing specific regulatory investigation costs. Vouch AI offers affirmative coverage for AI errors and omissions, bias and discrimination claims, intellectual property infringement, and regulatory investigations. The AI liability insurance market was valued at approximately $2 billion in cyber liability coverage alone in 2024, projected to reach $9 billion by 2035 (WiseGuy Reports, 2026).

Insurance markets function as governance mechanisms because insurers price risk based on the quality of an organization's risk management and accountability architecture. An organization with documented domain bounding, HITL interrupt checkpoints, immutable audit trails, and board-level AI oversight — the mechanisms specified in Paper 3 — presents a different risk profile to an insurer than an organization deploying AI without these mechanisms. Differential pricing translates governance quality into a private financial incentive: organizations with stronger accountability architecture pay lower premiums, creating a direct return on governance investment that accrues specifically to the investing organization rather than to the system as a whole. This is the selective incentive structure Olson identifies as capable of sustaining collective action in large groups.

The insurance mechanism is currently underdeveloped as a governance driver because underwriters have limited capacity to assess AI governance quality and because the actuarial history for AI liability claims is too thin to support robust premium differentiation. As the Moffatt v. Air Canada precedent (2024) and subsequent rulings establish liability patterns for AI system failures, and as frameworks like ISO/IEC 42001 provide standardized criteria for governance quality assessment, the insurance market's capacity to price governance differentially will improve. The policy implication is that mandatory AI insurance requirements — analogous to mandatory automobile or professional liability insurance in other domains — would accelerate this dynamic by creating a

universal market for AI liability coverage and forcing the development of actuarial frameworks for governance quality assessment.

### 5.2 Certification Markets

ISO/IEC 42001:2023, published in December 2023, is the first international standard for AI management systems. It specifies requirements for establishing, implementing, maintaining, and continually improving an AI management system within an organization, with a focus on transparency, fairness, accountability, and risk mitigation (ISO, 2023). The standard aligns with the NIST AI RMF and supports compliance with the EU AI Act. Microsoft has achieved ISO/IEC 42001 certification, with independent third-party audit reports available through its Service Trust Portal (Microsoft, 2026). The number of organizations achieving ISO certification broadly increased by 20% in 2024 compared to 2023, and ISO/IEC 42001 adoption is expected to grow steadily as AI becomes mainstream (Prompt Security, 2025).

Certification markets function as governance mechanisms because they create a verifiable, third-party-audited signal of governance quality that organizations can communicate to customers, regulators, partners, and investors. An ISO/IEC 42001 certificate is a selective incentive: it is available only to organizations that meet the standard's requirements, and its value as a market signal accrues specifically to certified organizations rather than to the market as a whole. Organizations that achieve certification can compete on governance quality as a differentiator, creating a private return on accountability investment.

The certification mechanism complements the insurance mechanism: insurers can use ISO/IEC 42001 certification as a governance quality proxy for premium assessment, creating a direct financial pathway from certification to lower insurance costs. This integration of certification and insurance markets into a mutually reinforcing governance incentive structure is the market mechanism analog of the regulatory compliance floor established by mandatory standards.

### 5.3 Disclosure Requirements

A third market mechanism is mandatory disclosure of AI governance quality and AI system usage in material business functions. Disclosure requirements are a lighter-touch regulatory intervention than mandatory standards: they do not specify minimum governance levels but require organizations to make their governance practices visible to market participants, who can then reward or penalize governance quality through investment, procurement, and partnership decisions.

The McKinsey (2025) finding that only 39% of Fortune 100 companies disclosed any form of board-level AI oversight as of 2024 — despite 88% using AI in core business functions — illustrates the disclosure gap. An organization that uses AI in consequential organizational decisions without disclosing this to investors and customers is not competing on governance quality; it is concealing the absence of it. Mandatory AI usage and governance disclosure requirements would eliminate this concealment advantage, forcing governance quality into the competitive calculus of market participants.

## 6. Mechanism 3: Polycentric Coordination and Shared Infrastructure

The third mechanism for addressing the collective action problem is coordination that creates shared governance infrastructure, reducing the per-organization cost of accountability architecture by pooling investment across industry participants. Ostrom's (1990) work on governing common-pool resources demonstrated that polycentric arrangements — multiple overlapping governance mechanisms operating at different levels — are more robust than either centralized top-down regulation or uncoordinated individual action. Recent research applying Ostrom's framework to global AI governance suggests that a polycentric multilevel arrangement will be more effective than any single centralized mechanism (Springer, 2025).

Several forms of polycentric coordination are emerging in the AI governance landscape. The network of AI Safety Institutes (AISIs) established at the Seoul Summit, now including ten countries with the UK AI Safety Institute as the inaugural member, creates shared testing and evaluation infrastructure that individual organizations and governments can access rather than build independently. This shared infrastructure lowers the per-organization cost of pre-deployment safety evaluation, one of the most technically demanding components of accountability architecture.

To function as an effective polycentric governance mechanism rather than a symbolic institutional network, the AISI network requires operational integration with the market and regulatory mechanisms described in Sections 4 and 5. Three specific integration pathways would convert AISI testing capacity into active economic incentives. First, AISI evaluation as a fast-track compliance safe harbor: organizations that submit AI systems operating at AAMM Level 2 or above to a recognized AISI for pre-deployment evaluation and receive a clearance determination could be granted a streamlined EU AI Act conformity assessment pathway, reducing the compliance cost and timeline for organizations that invest in independent pre-deployment testing. This creates a direct financial return on AISI engagement that does not depend on regulatory goodwill alone. Second, AISI clearance as an insurance premium tier anchor: insurers offering AI liability coverage could formally recognize AISI evaluation outcomes as a governance quality signal, assigning verified AISI-evaluated systems to lower premium tiers than self-certified systems. This integrates the AISI network into the insurance market mechanism and creates a second private financial return on testing investment. Third, AISI evaluation as a public procurement requirement: government procurement of AI systems at AAMM Level 2 and above could require AISI evaluation as a condition of contract eligibility, creating a mandatory market channel for AISI services that sustains the network financially and establishes a procurement-level governance floor independent of the general regulatory framework.

Industry consortia for shared governance standards — analogous to the Payment Card Industry Data Security Standard (PCI DSS) model in financial services, where industry participants collectively define and maintain security standards that apply to all members — represent another form of polycentric coordination. The advantage of the consortium model is that governance standards developed

by industry participants have higher practical specificity and implementability than standards developed exclusively by regulators, because industry participants understand the technical and operational constraints that shape implementation.

Shared audit infrastructure represents a third form of polycentric coordination. The development of standardized AI audit methodologies, accessible to third-party auditors and applicable across organizations and sectors, reduces the cost and increases the credibility of independent algorithmic auditing. A market for AI audit services analogous to the financial audit market — with recognized methodologies, accredited auditors, and standardized reporting — would substantially lower the per-organization cost of the independent auditing requirements identified in Paper 1 as a core governance mechanism.

The critical design principle for polycentric coordination mechanisms, derived from Ostrom's analysis, is that they must have clear boundaries (who is subject to the governance arrangement), rules matched to local conditions (governance requirements that reflect the specific risks and contexts of different AI applications), collective choice arrangements (mechanisms for participants to modify the governance rules over time), monitoring (independent verification of compliance), and graduated sanctions (consequences that scale with the severity and frequency of governance failures). Coordination mechanisms that lack these design features tend to degrade into lowest-common-denominator standards or to be captured by participants with the strongest incentives to minimize governance costs.

## 7. The Three Mechanisms in Interaction

| Mechanism | Olson Type | Addresses | Primary Limitation |
|---|---|---|---|
| Mandatory standards (EU AI Act) | Coercion | Free-rider advantage of non-compliance; levels competitive floor | Geographic scope; enforcement capacity; innovation friction |
| Insurance markets | Selective incentive (financial) | Private return on governance investment; actuarial accountability | Actuarial immaturity; requires liability precedent to price accurately |
| Certification markets (ISO 42001) | Selective incentive (reputational) | Verifiable governance signal; market differentiation | Voluntary adoption; signal inflation risk if standards diluted |
| Disclosure requirements | Selective incentive (reputational) | Concealment advantage of non-disclosure; market visibility | Requires investor/customer capacity to evaluate disclosures |
| Polycentric coordination (consortia, AISI network) | Cost reduction | Per-organization governance cost; shared testing infrastructure | Capture risk; lowest-common-denominator dynamics without Ostrom design principles |

*Table 1. Governance mechanisms, their structural type, and their primary limitations. No single mechanism is sufficient; the governance ecosystem requires all three to address the collective action problem at different levels.*

The three mechanisms are complementary rather than substitutes. Mandatory standards establish the floor below which no market participant can compete on governance cost. Market mechanisms create incentives to exceed that floor by making governance quality differentially rewarded. Polycentric coordination reduces the cost of meeting and exceeding the floor by sharing governance infrastructure. Each mechanism addresses a different dimension of the collective action problem, and each has limitations that the others partially mitigate.

The temporal sequencing of the mechanisms also matters. Voluntary commitments and polycentric coordination tend to precede mandatory standards historically, because they establish norms and technical feasibility that make mandatory requirements politically viable. The Seoul commitments preceded and informed elements of the EU AI Act's requirements. The NIST AI RMF, developed through a voluntary consensus process, provides technical content that mandatory standards can reference. Market mechanisms tend to develop in parallel with or slightly behind regulatory requirements, as insurers and certification bodies develop the actuarial and assessment infrastructure needed to operationalize governance quality as a market variable.

## 8. The Conditions for Self-Sustaining Governance Investment

The governance premium evidence from Paper 2 — that organizations with AI-savvy boards outperform their industry average by 10.9 percentage points in ROE (Weill, Woerner & Banner, 2025) — raises a question the preceding analysis does not fully address: if governance investment produces this magnitude of financial advantage, why does the free-rider problem persist at all? Why don't organizations invest in governance simply because it pays?

The answer is that the governance premium is a long-run, population-level finding. It describes the average performance difference between AI-savvy and non-AI-savvy board companies across a large sample of organizations over time. The individual organization facing a specific deployment decision operates on a shorter time horizon and faces a different calculation: the governance cost is certain and immediate; the governance premium is probabilistic and deferred. Under standard discounting, rational decision-makers will underweight the deferred premium relative to the immediate cost, particularly when competitive pressure creates urgency around deployment speed.

There are, however, conditions under which governance investment becomes self-sustaining as a competitive strategy rather than a regulatory burden. Understanding these conditions is important because they identify where governance premium dynamics can reinforce regulatory and market mechanisms rather than working against them.

Governance quality is observable. The governance premium can only function as a competitive differentiator when customers, investors, and partners can

observe governance quality and reward it. Disclosure requirements and certification schemes — particularly ISO/IEC 42001 certification with independent third-party verification — are the primary mechanisms for making governance quality observable. Organizations that achieve and disclose certification can capture the governance premium as a market signal; those that do not cannot differentiate on governance quality regardless of their actual practices.

Governance failures are attributable. The Moffatt v. Air Canada (2024) ruling established that organizations are liable for AI system failures in customer-facing contexts. As liability precedent develops across jurisdictions and domains, governance failures become increasingly attributable to specific organizations rather than diffuse across the system. When an organization bears the full cost of its governance failures — through litigation, regulatory penalties, and reputational damage — the private incentive to invest in governance increases correspondingly. The insurance market mechanism reinforces this: differential premium pricing for governance quality translates the attributable cost of governance failure into a continuous financial incentive for governance investment.

The organizational time horizon is long enough. The governance premium accrues over time; the governance cost is immediate. Organizations with short planning horizons — those facing quarterly earnings pressure, activist investors with short-term return expectations, or governance structures that incentivize near-term performance at the expense of long-term resilience — will systematically underweight the governance premium regardless of its magnitude. This suggests that corporate governance quality — the internal accountability structures of the deploying organization, not just its AI governance specifically — is a prerequisite for AI governance investment to be sustained without external regulatory compulsion.

These three conditions point toward a coherent policy agenda: mandatory disclosure makes governance quality observable; liability law and mandatory insurance make governance failures attributable; and corporate governance reform — including the board-level AI oversight requirements established by the EU AI Act and recommended by the MIT CISR research — extends organizational time horizons. When all three conditions are met, the governance premium becomes self-sustaining: organizations invest in governance because it is competitively rational to do so, not because they are compelled to.

## 9. Implications for the AAMM

The collective action analysis of this paper has direct implications for the AAMM framework developed across this series. The AAMM classifies AI systems by their level of decision-making authority, with governance requirements escalating from Level 0 through Level 3 as specified in Paper 3. The policy analysis of this paper suggests that the structural conditions for governance investment also vary by AAMM level.

At AAMM Levels 0 and 1, the collective action problem is relatively tractable. The consequences of governance failures are limited in scope, liability attribution is relatively straightforward (as the Air Canada case demonstrates), and the cost of the required governance mechanisms — transparency, output guardrails, usage logging — is low relative to the deployment cost of the AI system. Market mechanisms and reputational incentives may be sufficient to sustain governance investment at these levels without mandatory regulatory requirements.

At AAMM Level 2, the collective action problem becomes acute. The governance requirements — domain bounding, HITL interrupt architecture, immutable audit trails — are substantially more expensive to implement, the competitive pressure to deploy autonomously without these mechanisms is correspondingly stronger, and the consequences of governance failure are more severe. Mandatory standards become necessary at this level, which is why the EU AI Act's high-risk system requirements — covering autonomous AI systems in consequential decision domains — align closely with the Level 2 governance requirements specified in Paper 3.

At AAMM Levels 3 and above, the collective action problem reaches its most severe form. The governance requirements are most demanding, the competitive pressure for ungoverned deployment is strongest (because the strategic framing function of Level 3 systems provides significant competitive advantage to early adopters), and the consequences of governance failure — including the accountability void and fiduciary incompatibility identified in Papers 1 and 3 — are potentially systemic rather than organization-specific. No currently deployed regulatory framework adequately addresses Level 3 governance requirements, and no market mechanism yet provides the selective incentives required to sustain Level 3 governance investment voluntarily. This is the governance frontier: the level at which the structural conditions for accountable AI authority remain to be established.

## 10. Limitations

The collective action diagnosis is structural, not empirical. This paper applies Olson's theoretical framework to the AI governance context by analogy. It does not present empirical evidence that organizations are consciously free-riding on competitors' governance investment, or that the governance-deployment gap is in fact produced by the free-rider dynamic rather than other causes (ignorance, organizational inertia, technical difficulty). The diagnosis is plausible and consistent with the available evidence, but it has not been tested against alternative explanations.

The policy prescriptions are context-dependent. The effectiveness of mandatory standards, market mechanisms, and polycentric coordination depends on institutional context that varies substantially across jurisdictions, sectors, and organizational types. The EU AI Act analysis draws on a specific regulatory regime; other jurisdictions face different political and institutional constraints. The insurance and certification market analyses are based on an early and rapidly evolving market; the mechanisms described may develop differently from what the current trajectory suggests.

The governance premium evidence is correlational. As noted in Paper 2, the MIT CISR finding that AI-savvy boards outperform by 10.9 percentage points in ROE establishes association rather than causation. The analysis in Section 8, which builds on this finding to identify conditions for self-sustaining governance investment, inherits this limitation.

The international dimension is underspecified. The inter-jurisdictional race-to-the-bottom dynamic identified in Section 2 is acknowledged but not fully analyzed. A complete treatment of the competitive deployment dilemma at the international level would require engagement with international political economy and trade law literature that is beyond the scope of this paper.

## 11. Conclusion

The governance gap documented across this series is not an anomaly of organizational irresponsibility or regulatory failure. It is the predictable outcome of a collective action structure in which the costs of governance are private and the benefits are shared. Understanding it as such reframes the policy challenge: the question is not how to persuade organizations to govern AI responsibly, but how to restructure the incentive environment so that responsible governance is the individually rational choice.

The three mechanisms examined in this paper — mandatory standards, market mechanisms, and polycentric coordination — address the collective action problem at different levels and through different channels. None is sufficient alone. Mandatory standards establish the governance floor but cannot reach organizations outside their jurisdictional scope. Market mechanisms create differentiated rewards for governance investment but require observable governance quality and attributable governance failures to function effectively. Polycentric coordination reduces governance costs but is vulnerable to capture and lowest-common-denominator dynamics without robust design.

The conditions under which governance investment becomes self-sustaining — observable governance quality, attributable governance failures, and organizational time horizons long enough to capture the governance premium — suggest a coherent policy agenda: disclosure requirements, liability law development, mandatory insurance, and corporate governance reform that extends board-level AI oversight. When these conditions are met, the governance premium identified by MIT CISR (10.9 percentage points in ROE) creates a private financial incentive for governance investment that does not require ongoing regulatory compulsion.

Across four papers, this series has established what the AI governance challenge is, where it manifests in practice, what technical architecture it requires, and why market forces alone will not deliver it without structural intervention. The argument has moved from framework to evidence to engineering to policy. The practical work of building the institutional infrastructure — the regulatory frameworks, the insurance markets, the certification schemes, the coordination mechanisms — that makes accountable AI authority economically rational rather than merely admirable remains the central challenge for organizations,

regulators, and civil society alike.

---

## References


Computer Weekly. (2024, May 22). AI Seoul Summit: 16 AI firms make voluntary safety commitments. https://www.computerweekly.com/news/366585914/AI-Seoul-Summit-16-AI-firms-make-voluntary-safety-commitments

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council (EU AI Act). Official Journal of the European Union. https://artificialintelligenceact.eu/

Huseby, A. (2025a). Governance in the age of AI leadership: From advisory systems to organizational authority. *Paper 1 in a Series on AI Governance and Organizational Authority*. Zenodo. https://doi.org/10.5281/zenodo.20330936

Huseby, A. (2026a). The AAMM in practice: Classifying AI decision-making authority across enterprise deployments. *Paper 2 in a Series on AI Governance and Organizational Authority*. Zenodo.

Huseby, A. (2026b). Designing accountability in: Technical architecture and human oversight across AAMM levels. *Paper 3 in a Series on AI Governance and Organizational Authority*. Zenodo.

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

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360.

McCarthy Tétrault LLP. (2024). Moffatt v. Air Canada, 2024 BCCRT 149. https://www.mccarthy.ca/en/insights/blogs/techlex/moffatt-v-air-canada-misrepresentation-ai-chatbot

McKinsey & Company. (2025, December 4). The AI reckoning: How boards can evolve. https://www.mckinsey.com/capabilities/mckinsey-technology/our-insights/the-ai-reckoning-how-boards-can-evolve

Microsoft. (2026). ISO/IEC 42001:2023 artificial intelligence management system standards. Microsoft Learn. https://learn.microsoft.com/en-us/compliance/regulatory/offering-iso-42001

NIST. (2023). *Artificial Intelligence Risk Management Framework* (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups.* Harvard University Press.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action.* Cambridge University Press.

Prompt Security. (2025). Understanding ISO/IEC 42001 for AI management systems. https://prompt.security/blog/understanding-the-iso-iec-42001

Russell, S. J. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* Viking.

Secure Privacy. (2026). EU AI Act 2026: Key compliance requirements for enterprises. https://secureprivacy.ai/blog/eu-ai-act-2026-compliance

Springer. (2025). In search of a global governance mechanism for artificial intelligence: A collective action perspective. *Global Public Policy and Governance.* https://doi.org/10.1007/s43508-025-00113-z

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, 20(3), 571–610.

Tabor, F. (2026, February 4). AI insurance underwriting in 2026: Pricing risk, liability, and coverage in the age of artificial intelligence. https://www.francescatabor.com/articles/2026/2/4/ai-insurance-underwriting-in-2025-pricing-risk-liability-and-coverage-in-the-age-of-artificial-intelligence

Weill, P., Woerner, S. L., & Banner, J. (2025, December 8). AI-savvy boards drive superior performance. *MIT Sloan Management Review.* Based on MIT Center for Information Systems Research Briefing released March 20, 2025. https://sloanreview.mit.edu/article/ai-savvy-boards-drive-superior-performance/

Latham & Watkins. (2026, May). EU AI Act update: EU resolves to change rules and extend deadlines. https://www.lw.com/en/insights/ai-act-update-eu-resolves-to-change-rules-and-extend-deadlines

WiseGuy Reports. (2026). *Artificial Intelligence AI Liability Insurance Market 2035.* https://www.wiseguyreports.com/reports/artificial-intelligence-ai-liability-insurance-market

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2020). Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10), 3267–3297.

Dou, W. W., et al. (2025). Financial stability implications of generative AI. Federal Reserve Board Finance and Economics Discussion Series 2025-090. https://www.federalreserve.gov/econres/feds/files/2025090pap.pdf

European Commission. (2026). Carbon Border Adjustment Mechanism: Definitive regime implementation. Taxation and Customs Union. https://taxation-customs.ec.europa.eu/carbon-border-adjustment-mechanism_en

Global Treasurer. (2025, February 25). AI speed presents risks to financial markets. https://www.theglobaltreasurer.com/2025/02/25/ai-speed-presents-risks-to-financial-markets/

IMF. (2024). Global Financial Stability Report, Chapter 3: Financial Stability Implications of AI in Financial Markets. International Monetary Fund, October 2024.

Lin, P. (2025, November 25). AI and the future of market manipulation. *The Regulatory Review.* https://www.theregreview.org/2025/11/25/smith-ai-and-the-future-of-market-manipulation/

US Treasury. (2024, December). *Report on Artificial Intelligence in Financial Services.* United States Department of the Treasury.

---