Governance in the Age of AI Leadership: From Advisory Systems to Organizational Authority

(AI as CEO — Mapping the Trajectory, Limits, and Governance Imperatives of Autonomous AI Leadership)

Paper 1 in a Series on AI Governance and Organizational Authority

---

Alexander Huseby
Founder, Cognitive Liberty Institute
Oslo, Norway — alexander@coglib.no
2025

Abstract

The question of whether artificial intelligence can occupy the role of Chief Executive Officer has shifted from theoretical provocation to live experiment. In October 2025, Kazakhstan's sovereign wealth fund Samruk-Kazyna formally elected SKAI — an AI system — as an independent board director with claimed voting rights, though the legal enforceability of that authority remains disputed under Kazakh law. Researchers at the Wharton Mack Institute and INSEAD's Center for Corporate Governance have begun formally comparing AI and human boards on governance criteria. The Dataiku Global AI Confessions Report (Harris Poll, March 2025), surveying over 500 CEOs across the US, UK, France, and Germany, found that 94% believe AI could offer equal or better counsel than at least one of their current board members. This paper examines what these developments actually mean: what AI can do in an executive capacity, where it demonstrably cannot, and what governance architecture is required if organizations are to move from AI-assisted leadership to AI-principal leadership responsibly. Drawing on principal-agent theory (Jensen & Meckling, 1976), organizational legitimacy theory (Suchman, 1995), and institutional governance frameworks (Ostrom, 1990), it introduces a six-level AI Authority Maturity Model and proposes governance as a form of strategic infrastructure rather than a compliance cost.

Keywords: AI leadership, corporate governance, autonomous AI agents, AI accountability, principal-agent theory, organizational legitimacy, CEO decision-making, agentic AI, AI ethics, governance maturity

---

1. The Question Has Changed

For most of the last decade, "could AI be a CEO?" was a thought experiment — useful for probing assumptions about leadership, consciousness, and organizational dynamics, but not a practical question. That changed around 2024.

The signals are accumulating. Deep Knowledge Ventures appointed an algorithmic system called VITAL as a board observer with advisory participation as early as 2014, though it functioned more as a data processing tool rather than a genuine decision-maker. A decade later, the capabilities have changed substantially. In October 2025, Kazakhstan's sovereign wealth fund Samruk-Kazyna formally elected SKAI — a large language model-based system — as an independent board director, notifying the Kazakhstan Stock Exchange of the appointment. The fund described SKAI as holding voting rights on strategic and financial decisions. However, a Kazakh member of parliament subsequently noted that the country's AI law explicitly states that AI is not a legal subject and therefore cannot hold office, sign documents, or vote; the legal enforceability of SKAI's board authority remains unresolved. What is not in dispute is that a national sovereign wealth fund formally elected an AI system to its board — a development without precedent in corporate governance.

Separately, researchers at the Wharton Mack Institute and INSEAD's Center for Corporate Governance ran a controlled experiment in 2025 pitting an AI board against a human board deliberating the same business case. The AI board, built as a multi-agent simulation, processed materials instantly, followed governance protocols without deviation, and was evaluated by independent assessors across eight governance criteria (Yakubovich & Shekshnia, Harvard Business Review, November 2025).

Meanwhile, agentic AI systems — capable of autonomous multi-step reasoning and action — have moved from research prototypes to enterprise deployment. JPMorgan's COIN system reduced hundreds of thousands of hours of annual legal document review to highly automated workflows completed in minutes. Salesforce Einstein operates as an embedded AI decision-support system used across thousands of organizations. Cognition's Devin, launched in 2024, was designed to autonomously plan and execute multi-step engineering tasks.

More than 88% of organizations now use AI in at least one core business function (McKinsey, 2025). Yet as of 2024, only 39% of Fortune 100 companies disclosed any form of board-level AI oversight (McKinsey, 2025). The gap between deployment and governance is the central problem this paper addresses.

A definitional note: this paper uses "AI" to refer primarily to large language model-based generative systems and agentic AI — autonomous multi-agent systems capable of multi-step planning and execution. Predictive AI and narrow analytics tools represent an earlier and largely distinct category. For purposes of this paper, *AI-principal leadership* refers to AI systems possessing formal or de facto authority over strategic organizational decisions beyond purely advisory functions.

2. A Framework for AI Authority in Organizations

The literature on AI in organizations frequently conflates fundamentally different categories: AI as a data tool, AI as an advisor, AI as an autonomous executor, and AI as a governance participant. These distinctions carry significant legal, ethical, and organizational consequences. To clarify the trajectory under discussion, this paper introduces the AI Authority Maturity Model (AAMM), a six-level taxonomy describing the progression of AI decision-making authority in organizational contexts.

Level	Role	Human Oversight	Example
0	Analytics tool	Full — human acts on output	BI dashboards, reporting
1	AI advisor / copilot	Human retains all decisions	Salesforce Einstein, Copilot
2	Autonomous executor (bounded domain)	Human approval gates	COIN, supply chain AI
3	Strategic recommender	Human ratification required	AI board preparation tools
4	Governance participant	Shared or disputed authority	SKAI, Kazakhstan 2025
5	Principal executive	Exceptional / minimal oversight	Hypothetical — not yet realized

Table 1. AI Authority Maturity Model (AAMM). Each level represents a qualitatively distinct relationship between AI systems and organizational decision-making authority.

Most currently deployed enterprise AI systems operate at Levels 0–2. The Wharton-INSEAD experiment operated at Level 3–4. The Kazakhstan SKAI appointment sits at Level 4, with the legal ambiguity around its actual voting power reflecting precisely the institutional gap between claimed and enforceable authority. Level 5 remains hypothetical, though the trajectory of the preceding levels points toward it.

The principal-agent problem — first formalized by Jensen and Meckling (1976) — describes the misalignment of incentives between a principal (shareholder, board) and an agent (manager) acting on their behalf. AI leadership introduces a compounded version of this problem: the AI system is simultaneously an agent of its deploying organization and a product of its developers' design choices. Unlike human agents, AI systems cannot be held legally responsible for the decisions they make, cannot bear reputational consequences, and have no interests of their own to align. The alignment mechanisms that classical principal-agent theory relies upon — compensation structures, reputation, career incentives — have no purchase. New alignment mechanisms are required.

3. What AI Does Well in an Executive Context

It is necessary to be precise about capability. Three domains stand out where AI demonstrably outperforms human executives in bounded conditions. It bears noting that most currently deployed systems remain domain-specific,

human-supervised, and organizationally constrained — the following reflects demonstrated performance within those bounds, not general executive agency.

Data processing and pattern recognition at scale. Human executives make decisions under severe information constraints, relying on filtered summaries and trusted intermediaries. AI systems face none of these constraints. BlackRock's Aladdin platform monitors risk across more than $9 trillion in assets by processing variables no human team could hold simultaneously. This is not a marginal efficiency gain — it represents a qualitatively different class of decision input.

Consistency and freedom from self-serving bias. Human CEOs are subject to well-documented cognitive distortions: overconfidence, loss aversion, in-group favoritism, and motivated reasoning that protects their tenure. AI systems do not have careers to protect, boards to impress, or compensation structures that incentivize short-term earnings manipulation. This theoretical objectivity is significant — provided the system's training data and objective function are themselves unbiased, a substantial caveat addressed in Section 4.

Speed and continuous availability. Markets move continuously, crises arrive without schedule, and competitive intelligence degrades rapidly. AI systems operate without the biological constraints limiting human executives. In domains where decision latency is a competitive variable — financial services, supply chain, cybersecurity — continuous availability is a structural advantage.

These capabilities explain why the trajectory toward expanded AI decision-making authority is economically rational. The pressure to expand AI authority is not ideological. It is competitive.

4. The Limits That Actually Matter

The case for AI limitations in leadership is frequently made poorly — anthropomorphic arguments about AI lacking "heart" or "vision" that are emotionally resonant but analytically weak. The genuine limits cluster around four specific problems.

The objective function problem. AI systems optimize for what they are told to optimize for. Corporate leadership requires navigating objectives that are poorly specified, mutually contradictory, and dynamically shifting: shareholder returns versus employee welfare versus long-term brand integrity versus regulatory compliance versus community relationships. No current AI system has a principled way to adjudicate between these when they conflict, because there is no principled way to encode the full complexity of human values into a single objective function. This is not a temporary technical limitation. It is a deep unsolved problem in AI alignment (Russell, 2019).

Training data encodes the past. AI systems learn from historical data, inheriting historical biases and historical definitions of good performance. A system trained on decades of Fortune 500 outcomes would likely reproduce that population's patterns — including systematic underinvestment in long-term resilience, bias toward financial metrics over social ones, and demographic homogeneity in

leadership selection. The appearance of objectivity can mask a conservative, backward-looking disposition.

Unprecedented situations. The most consequential executive decisions are precisely those without historical precedent. AI systems degrade in exactly these conditions — not because they lack processing power, but because their pattern-matching requires patterns to match against. Human executives can reason from first principles in ways current AI systems cannot reliably replicate.

Accountability has no home. When a human CEO makes a catastrophic decision, a clear accountability structure exists: board oversight, shareholder litigation, personal reputational consequences, and in extreme cases criminal liability. When an AI system makes a catastrophic decision, accountability diffuses across developers, the deploying organization, the approving board, and the training data. Under corporate law, fiduciary duty requires a human agent capable of bearing legal responsibility. AI systems cannot satisfy that requirement in any jurisdiction as of 2025. The EU AI Act represents the most serious attempt to address this, but stops well short of attributing legal personhood or liability to AI systems.

4.1 Why AI Principal Leadership May Never Fully Emerge

Beyond the technical limitations above, there are structural reasons why full AI principal leadership may remain permanently constrained. Drawing on Suchman's (1995) framework, organizational legitimacy — the social license that allows a CEO to lead — derives from pragmatic, moral, and cognitive sources. AI systems face deficits on all three dimensions: they cannot be trusted in the same way as human agents, they cannot bear moral responsibility, and their authority as organizational principals has not achieved the taken-for-granted status that cognitive legitimacy requires.

Labor resistance, political resistance, and the fundamental incompatibility between AI decision-making and fiduciary law are structural constraints, not merely transitional ones. Even if AI capability eventually surpasses human performance across all measurable executive functions, whether AI should hold principal authority is a question of institutional design, not capability. Trust asymmetry — stakeholders' greater willingness to accept human error than AI error of the same magnitude — represents a durable social fact that governance frameworks must accommodate rather than assume away.

5. The Governance Gap

McKinsey's 2025 assessment found that fewer than 25% of companies have board-approved, structured AI policies — despite more than 88% using AI in core business functions. This governance gap — deployment dramatically outpacing the institutional frameworks needed to make it accountable — is the central organizational risk of this moment.

The absence of governance creates perverse incentives. Organizations under competitive pressure to deploy AI will do so without adequate oversight if oversight is not mandated. The rational individual actor — a CEO facing quarterly

earnings pressure — has strong incentives to expand AI authority quickly and weak incentives to invest in governance infrastructure that slows deployment and adds cost.

Governance designed for human principals does not transfer to AI principals. Board oversight, fiduciary duty, duty of care — these constructs assume a human agent whose behavior can be observed, whose motivations can be interrogated, and who can be held personally accountable. The distinction between AI assistance, AI recommendation, AI autonomy, AI authority, and AI accountability matters enormously in legal and governance terms, yet most organizations treat these as a single undifferentiated category. New governance constructs are needed, not adaptations of existing ones.

Ostrom's (1990) work on governing common-pool resources demonstrates that effective institutional governance requires clearly defined boundaries, rules matched to local conditions, collective choice arrangements, monitoring, graduated sanctions, conflict resolution mechanisms, and recognition of rights to organize. These design principles translate with surprising fidelity to the AI governance context: the challenge of preventing individual actors from exploiting a shared resource — in this case, the trust infrastructure of organizational decision-making — without adequate collective governance is structurally similar.

The window for establishing governance norms is narrowing. Once AI systems are embedded in strategic decision-making at scale, the organizational dependencies and sunk costs make rolling back their authority extremely difficult. Governance frameworks established before widespread deployment have significant leverage; those established after have much less.

5.1 Concrete Governance Mechanisms

What would adequate governance actually require? The following mechanisms represent the minimum viable governance architecture for organizations operating AI systems at AAMM Levels 2 and above.

Mandatory audit trails. All AI-generated strategic recommendations must be logged with sufficient granularity to reconstruct the basis for any decision, analogous to existing requirements for board minutes.

Independent algorithmic auditing. Regular third-party audits of AI systems used in strategic functions, analogous to financial auditing. Making this mandatory rather than optional is the governance gap.

Board-level AI accountability. Explicit designation of a board member or committee responsible for AI oversight — not delegated to IT or legal, but treated as a governance function equivalent to audit or compensation.

Mandatory human veto authority. At AAMM Levels 3 and above, a defined human principal must retain explicit veto authority over AI-generated recommendations before they become binding decisions.

Red-teaming and adversarial testing. Structured processes that actively probe AI systems for failure modes before deployment in high-stakes decision contexts.

Model provenance disclosure. Organizations should disclose the AI systems used in material decisions, training data sources, and known limitations, analogous to existing material risk disclosures.

Liability insurance requirements. Requiring organizations to carry insurance for AI-driven decision failures creates market incentives for rigorous governance, as insurers will price risk based on governance quality.

6. The Hybrid Governance Transition

The most plausible medium-term outcome is not AI replacement of executive leadership, but hybrid governance architectures in which AI systems increasingly shape strategic options while humans retain legal and symbolic accountability. This is already the de facto situation in most sophisticated organizations: AI generates analysis, scenarios, and recommendations; humans ratify and bear responsibility for outcomes.

The progression toward hybrid governance is individually defensible at each step. Each step also makes the next step easier to justify. AAMM Level 1 normalizes AI input into decisions. Level 2 normalizes AI-initiated action within defined parameters. Level 3 normalizes AI framing of the decision space itself — which shapes outcomes even when humans nominally retain authority. Level 4 formalizes AI presence in the accountability structure. The cumulative effect of these incremental steps may exceed what any single step would have been approved on its own merits.

This incremental dynamic is precisely why governance frameworks established early — at Levels 1 and 2 — are substantially more effective than frameworks applied retrospectively at Levels 3 and 4. The organizational dependencies and reputational investments that accumulate around AI systems at each level make governance intervention progressively more costly.

The organizations that will navigate this transition well are not those that resist AI authority expansion, nor those that embrace it uncritically. They are those that treat governance as a competitive capability rather than a compliance cost. This is the paper's central normative claim: *in AI-intensive organizations, governance itself becomes a form of strategic infrastructure.* Robust AI governance — transparent, auditable, accountable — will increasingly be a condition of institutional trust, regulatory approval, long-term organizational legitimacy, and ultimately competitive advantage.

7. Limitations

The empirical evidence base remains sparse. Most cited cases are early-stage experiments or pilot programs rather than mature implementations with longitudinal performance data. The governance claims made here are therefore primarily theoretical and prospective.

The AI Authority Maturity Model introduced here is a conceptual framework, not an empirically validated instrument. It provides useful analytical clarity but has not been tested against a systematic sample of organizational AI deployments. Subsequent papers in this series address this gap through structured case analysis and technical specification.

Long-term trajectories are inherently speculative. The pace of capability development, regulatory response, and organizational adoption are all uncertain. The paper draws primarily on English-language sources from US and European contexts; governance challenges may differ substantially in other legal and cultural contexts.

8. Conclusion

The question is no longer whether AI will play a role in corporate leadership. It already does, at scale, across most significant organizations. The question is whether that role will expand with adequate accountability structures in place.

The evidence from 2024–2025 — the Samruk-Kazyna appointment, the Wharton-INSEAD board experiment, the Dataiku CEO survey — suggests the expansion is accelerating faster than the governance. That asymmetry is the central risk.

AI can process more information, maintain more consistency, and operate more continuously than any human executive. It cannot, in its current form, navigate genuine value conflicts, satisfy fiduciary requirements, or be held accountable in the ways that institutional legitimacy requires. Whether those limitations are permanent or transitional is an open question. What is not open is that building adequate governance frameworks requires treating them as if they matter.

The AAMM introduced here offers researchers and practitioners a shared vocabulary for discussing the progression of AI authority. Subsequent papers in this series apply the framework empirically, specify its technical requirements, and examine the policy conditions under which responsible deployment becomes economically rational.

The organizations, regulators, and researchers who engage seriously with that work now will be significantly better positioned than those who wait for a crisis to force the conversation.

References

Brynjolfsson, E., & McAfee, A. (2017). *Machine, Platform, Crowd: Harnessing Our Digital Future*. W.W. Norton & Company.

Conference Board. (2026). *C-Suite Outlook Survey: AI and Executive Strategy*. The Conference Board. https://www.conference-board.org

Davenport, T. H., & Kirby, J. (2015). Beyond automation: Strategies for remaining gainfully employed in an era of very smart machines. *Harvard Business Review*, 93(6), 58–66.

Dataiku. (2025, March 11). *Global AI Confessions Report: CEO Edition.* Conducted by The Harris Poll. https://blog.dataiku.com/confessions-from-500-global-ceos

Deloitte Insights. (2024). *Autonomous Generative AI Agents: Still Under Development.* Deloitte.

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer.

European Commission. (2021). *Proposal for a Regulation on Artificial Intelligence* (AI Act). European Commission.

Floridi, L. (2019). Soft ethics and the governance of the digital. *Philosophy & Technology*, *32*(1), 1–8.

Huang, M. H., & Rust, R. T. (2018). Artificial intelligence in service. *Journal of Service Research*, *21*(2), 155–172.

INSEAD Center for Corporate Governance & Wharton Mack Institute. (2025). Can AI boards outperform human ones? *Harvard Business Review*, November 2025.

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.

Kursiv Media. (2025, October 30). Kazakh wealth fund's AI board member has no real power, lawmaker says.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

McKinsey & Company. (2025). *The AI Reckoning: How Boards Can Evolve.* https://www.mckinsey.com/capabilities/mckinsey-technology/our-insights/the-ai-reckoning-how-boards-can-evolve

McKinsey & Company. (2026). *Developing Human Leadership in the Age of AI.* McKinsey Insights.

Mondaq / Dentons. (2025, November). AI joins Samruk-Kazyna's board of directors.

OECD. (2019). *Artificial Intelligence in Business and Management.* OECD Publishing.

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

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

Russell, S. J., & Norvig, P. (2022). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.

Samruk-Kazyna JSC. (2025, October 2). *First in the Region: Samruk-Kazyna Introduces AI-based Digital Board Member with Voting Rights.* https://sk.kz/press-center/news/78513/?lang=en

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

Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and AI systems. *Philosophical Transactions of the Royal Society A*, *376*(2133).

Yakubovich, V., & Shekshnia, S. (2025, November). Can AI boards outperform human ones? *Harvard Business Review.*

Zuboff, S. (2019). *The Age of Surveillance Capitalism.* PublicAffairs.