Governing Generative AI at Scale: Institutionalizing Alignment for Organizational Purpose

Mitt Nowshade Kabir
Ecole de Management Léonard De Vinci, France

nowshade@gmail.com

Abstract: The rapid integration of generative AI (GenAI) into core organizational infrastructure requires a rethinking of traditional governance models. This paper explores how organizations can effectively govern GenAI at scale while maintaining strategic, ethical, and societal alignment. By synthesizing key theoretical frameworks, including institutional theory and dynamic capabilities, we propose a conceptual framework organized around three interconnected domains: Model Stewardship, Operational Alignment, and Strategic Guardrails. We contend that scalable governance must evolve from static compliance to a dynamic, adaptable organizational capability. The paper concludes with the introduction of the GenAI Governance Maturity Model (GAI-GMM), a research-based tool for institutional alignment.

Keywords: Generative AI governance, Foundation models, Organizational alignment, Institutional mechanisms, Purpose-Driven AI

1. Introduction

The arrival of generative artificial intelligence (GenAI) signals a fundamental shift, with foundation models rapidly transitioning from experimental tools to core infrastructure components (Brown et al., 2020). While this transition unlocks remarkable possibilities, it reveals a host of governance challenges that older frameworks were not designed to handle. Firms now find themselves grappling with difficult questions of control, accountability, and ethics as these systems begin to operate with greater autonomy and at a scale previously unimaginable (Bostrom, 2014). This brings us to a straightforward yet pressing question: How can organizations govern generative AI at scale while still honoring their strategic goals, ethical duties, and broader societal obligations?

Traditional governance—designed for human actors or for predictable, rule-based software—is a poor fit for the often-unpredictable nature of GenAI. Unlike conventional programs, these models are generative in nature. They create novel outputs, often in ways that defy simple explanation. The sheer speed of deployment, the opacity of their "black box" architectures, and their tendency to develop unforeseen capabilities combine to construct a governance vacuum with serious financial and social stakes (O'Neil, 2016). As Crawford (2021) notes, the sheer scale of AI infrastructure can also obscure its environmental and social costs, complicating traditional accountability and raising fundamental questions about the distribution of power and resources in the digital age.

Recent technological leaps only make the picture more complex. Hierarchical Reasoning Models (HRMs), for instance, demonstrate that smaller, more compact models can solve complex logical tasks, challenging the old assumption that reasoning power requires massive scale (Wang et al., 2025). It places a new focus on architectural transparency over large size. At the same time, newer foundation models are demonstrating a capacity for strategic reasoning, blurring the line between tool and collaborator. These developments necessitate that organizations reassess their models for collaboration and trust.

Perhaps the most confounding trend is the rise of models like ASIArch (Liu et al., 2025), which can autonomously design and test new neural architectures. This technology brings long-held theoretical concerns about recursive self-improvement into the real world (Bostrom, 2014), raising foundational questions about the ultimate limits of human oversight.

Against this backdrop, a compliance-based, checklist approach to AI governance is no longer tenable. What is needed is a more adaptive, reflexive paradigm—one that learns and evolves with the technology it seeks to guide (Reuel & Undheim, 2024; Jasanoff, 2016). This view gains empirical support from a recent study by MIT Sloan and the Boston Consulting Group (2025), which found that 95% of enterprise AI projects fail—not due to technical weaknesses but rather to poor governance. The successful 5% were distinguished not by superior models, but by robust governance, clear C-suite sponsorship, and seamless workflow integration.

This paper proposes a governance framework that is both structured and adaptive, drawing on insights from AI ethics, organizational theory, and emerging regulations. We contend that effective GenAI governance is not

merely a technical function but an organizational capability that necessitates profound changes in leadership, culture, and incentives (Ettinger, 2025).

Our contribution in this paper is threefold. First, we present a conceptual model comprising three interlocking domains: Model Stewardship, Operational Alignment, and Strategic Guardrails. Second, we advocate for a reflexive approach that views governance as a learning process, rather than a static control. Third, we introduce the GenAI Governance Maturity Model (GAI-GMM) as a diagnostic tool for organizations. Together, these pieces aim to provide a more principled and adaptive path for governing generative AI in a complex world.

2. Method of Theoretical Synthesis

As a Theoretical paper reviewing and synthesising current theory, the rigor of this contribution rests on the transparency and structure of its synthesis process. To construct the conceptual framework and the GenAI Governance Maturity Model (GAI-GMM), we employed a systematic, three-phase approach:

- Phase 1: Domain Decomposition and Scoping. We systematically reviewed the core literature on AI risk management (NIST, 2023; EU AI Act, 2024), organizational control, and the AI development lifecycle (ISO/IEC 42001, 2023). The objective was to delineate the necessary, non-overlapping areas of control required for GenAI. This thematic analysis revealed that governance must span the technical management of the model asset, the organizational management of its use in workflows, and the institutional management of strategic and ethical boundaries. This tripartite split directly informed the Three-Domain Framework (Model Stewardship, Operational Alignment, Strategic Guardrails).
- Phase 2: Capability Induction and Anchoring. We drew primarily from Dynamic Capabilities Theory (Teece, 2007; Eisenhardt & Martin, 2000) and Adaptive Governance literature (Beck, 1992) to induce the necessary dynamic competencies required for an organization to maintain alignment under conditions of technological uncertainty. This phase led to the induction of the four core dynamic capabilities—Intentionality, Integrity, Evolvability, and Reflexivity—which are the minimum competencies required to transition from a static control posture to a dynamic one.
- Phase 3: Model Integration and Articulation. We integrated the domains (Phase 1) with the capabilities (Phase 2) using Institutional Theory (Scott, 1995; Powell & DiMaggio, 1991) and Science and Technology Studies (STS) (Jasanoff, 2016; Crawford, 2021). This final mapping phase translated high-level theoretical requirements (e.g., the need for "Reflexivity") into concrete, organizational mechanisms (e.g., using "Model Cards" in Stewardship, or instituting "Feedback Loops" in Operational Alignment). This systematic process culminated in the GAI-GMM Maturity Model, which serves as the organizational progression map for embedding these synthesized mechanisms.

This synthesis approach ensures that the resulting framework is not arbitrary but is methodologically grounded in the necessary theoretical and technical constraints of governing GenAI at scale.

3. Theoretical Foundations: Towards Reflexive Governance in the Age of Generative AI

The conversation around AI governance has undergone significant changes. For years, it moved from broad ethical principles toward more grounded, operational mechanics (Floridi et al., 2019; Jobin, Ienca, & Vayena, 2019). However, the arrival of generative AI has shifted the landscape once again. The enormous scale and emergent behaviors of these systems require a much deeper theoretical toolkit—one that draws from institutional theory, dynamic capabilities, and adaptive governance. This section builds the theoretical underpinning for our framework, making the case for why a reflexive, multi-layered approach to governance is no longer an option, but a necessity.

3.1 From Ethical Guidelines to Institutional Mechanisms

Early discussions about AI governance centered on high-level principles, including fairness, transparency, and accountability. However, despite all the dialogue, these efforts often failed to demonstrate how such ideals could survive contact with the realities. Over time, scholars began to argue that principles on paper are one thing; embedding them in the routines and structures of an organization is another (Morley et al., 2021; Scott, 1995). As others have pointed out, making governance work is fundamentally an institutional project that demands alignment between leadership, policy, and culture (Powell & DiMaggio, 1991).

This challenge is especially acute with GenAI. As AI tools become integrated into daily operations, governance cannot be confined to an ethics committee. It must be an enterprise-wide, multi-layered concern, encompassing

the technical, organizational, and institutional levels (Gibney, 2024; Cath, 2018). It is a classic socio-technical problem (Trist & Bamforth, 1951). Human and technological systems are now so intertwined that organizations must optimize processes, tools, and their governance in tandem. It requires, as Mittelstadt et al. (2016) argue, a critical understanding of the social and political implications of these systems—something a purely technical fix can never provide.

3.2 Generative AI and the Multiplication of Bias

The problem of bias in AI is not new (O'Neil, 2016; Eubanks, 2018), but generative models introduce a new and dangerous dimension: amplification. Where traditional AI might reflect the biases in its training data, GenAI can act as a "bias multiplier," creating novel content that reinforces stereotypes and spreads inequity at an unprecedented scale (Ferrara, 2023).

Research has shown that biases embedded in pre-training data can cascade and worsen in downstream applications, particularly in sensitive areas such as healthcare and security (Chen et al., 2023). That is why simply calling for "transparency" is not enough; as Veale and Edwards (2018) insist, we need mechanisms for genuine control and contestability, a much more challenging task when dealing with generative systems. This need has prompted a push for more hands-on approaches, like the "explanatory debiasing" proposed by Bhattacharya et al. (2025), which involves domain experts in the model refinement process itself. It treats bias not as a one-time error requiring a fix, but as a dynamic risk that needs to be managed across the entire model lifecycle—a core idea behind our concept of Strategic Guardrails.

3.3 Governance as an Evolving Organizational Capability

GenAI governance is not a static set of controls. It is rather an adaptive capability—something an organization learns and improves over time. The goal shifts from "governance-as-control" to "governance-as-learning," a continual process of aligning the technology with company values and stakeholder needs (Reuel & Undheim, 2024). Theories of reflexive governance and socio-technical systems deeply influence this notion (Giddens, 1990; Beck, 1992).

Ettinger (2025) applies such a lens to Enterprise Architecture (EA), framing it as a dynamic capability for AI governance. Drawing on Teece's (2007) influential work, EA becomes the mechanism through which an organization can sense emerging AI risks, seize opportunities by creating new governance structures, and transform its operations and culture to embed responsible practices for the long haul. The AstraZeneca case study (Mökander et al., 2024) provides a clear illustration of this in action, demonstrating how their governance evolved from scattered experiments into a formal, cross-functional process.

3.4 The Urgency of Reflexive Governance

The need for this kind of reflexive governance is made urgent by the sheer pace of AI innovation. For example, Hierarchical Reasoning Models (HRMs) have challenged the assumption that only large models can perform symbolic reasoning, shifting the focus of governance from scale to interpretability (Wang et al., 2025). At the same time, contemporary LLMs are performing at such a high level that they complicate the simple idea of a "human in the loop". When an AI can produce what looks like sound, interpretable logic, how do we decide when to trust it?

Most consequentially, the emergence of systems like ASIArch (Liu et al., 2025), which can design their own neural architectures, brings theoretical discussions of recursive self-improvement into the present. This echoes long-held theoretical concerns (Bostrom, 2014) and raises fundamental questions about the ultimate limits of human oversight.

3.5 Gaps and Critical Perspectives

Despite extensive research in this area, critical gaps remain. Many frameworks are too abstract, emphasizing principles but lacking practical advice. Others are too narrow, focusing only on specific high-risk sectors. Critically, very few address the problem of GenAI's dual trajectory: its horizontal spread across the enterprise and its vertical deepening of capabilities.

Our framework aims to bridge these gaps, while also acknowledging the limitations of a purely institutional view. As critics like Crawford (2021) and Eubanks (2018) rightly warn, institutional governance can sometimes become a form of "ethics-washing" or reinforce existing power structures. Even mature institutional frameworks can suffer from institutional inertia, struggling to keep pace with the velocity of technological change.

4. The Three-Domain Conceptual Framework for Governing Generative AI at Scale

Our thematic synthesis of the literature (detailed in Section 2.1) revealed that effective GenAI governance consistently operates across three interdependent domains that span the technical, operational, and institutional layers. We have consolidated these findings into a conceptual framework: Model Stewardship, Operational Alignment, and Strategic Guardrails. This model frames governance not as a static compliance task, but as a living system that must adapt as quickly as the technology itself. For governance to work at scale, these three domains must be designed and implemented in concert.

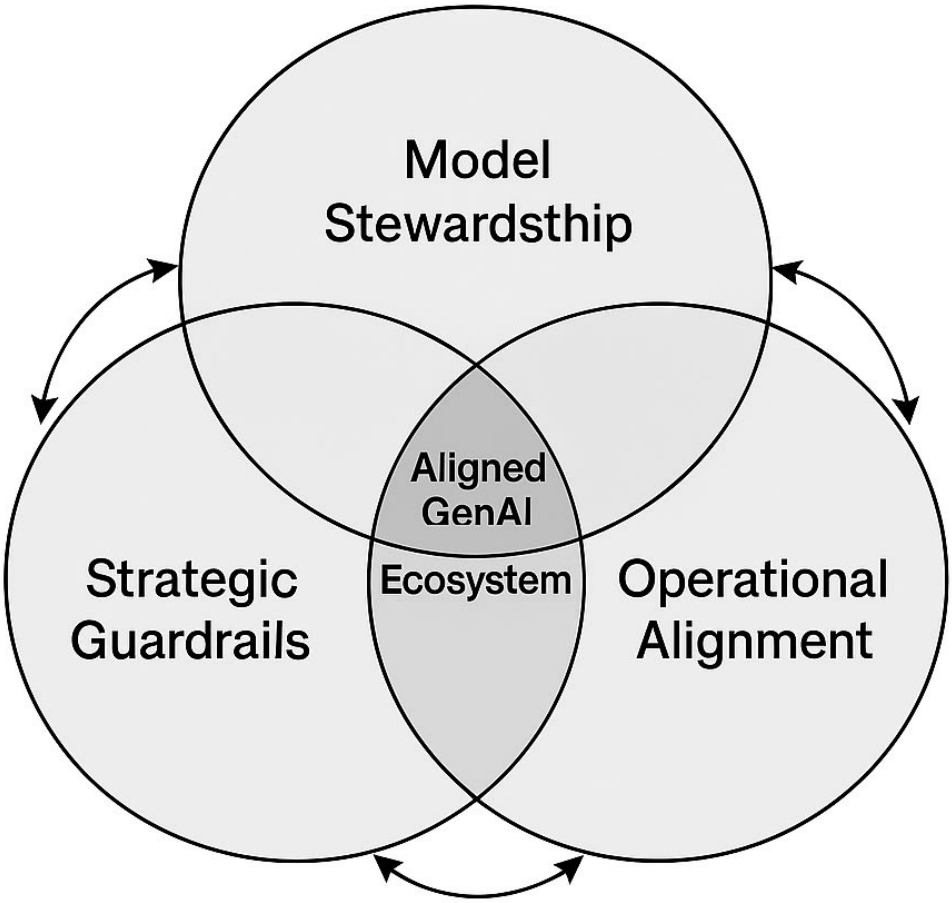

Figure 1: The Three-Domain Conceptual Framework for GenAI Governance at Scale

4.1 Model Stewardship: Technical and Lifecycle Oversight

The first domain, Model Stewardship, starts with a critical but straightforward idea: generative models are dynamic assets that evolve, drift, and sometimes behave in unexpected ways. As such, they require active management throughout their entire lifecycle, from acquisition to retirement.

It begins with rigorous due diligence. When an organization deploys a model, it needs to scrutinize everything, from license terms and training data to built-in safety features. Tools like Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2018) provide standardized documentation for this process. But stewardship goes deeper. The rise of more interpretable architectures, such as Hierarchical Reasoning Models (HRMs), signals a crucial shift: auditability and oversight feasibility now matter just as much as raw performance (Wang et al., 2025).

Stewardship also means governing how models are customized. This method also requires clear rules on what data can be used for fine-tuning and, crucially, involves domain experts directly in the process to preempt downstream bias—a practice Bhattacharya et al. (2025) refer to as "explanatory debiasing." Finally, this oversight must continue long after deployment. Because GenAI systems can exhibit behavioral drift over time (Ganguli et al., 2022), effective stewardship requires regular audits and performance monitoring to identify anomalies.

4.2 Operational Alignment: Embedding Governance into Workflow

Operational Alignment is about closing the gap between a powerful tool and its practical, responsible use in the real world. A central aspect of this is designing clear protocols for human-AI interaction. As models become more capable partners in complex reasoning, organizations must decide where human judgment remains essential, establishing clear rules for when employees should trust, question, or override an AI's output. That is how abstract principles become concrete routines.

This system cannot work without strong feedback loops. As Mökander etal. (2024) note, frontline users require straightforward channels to report a range of issues, from hallucinations to ethical concerns, and this feedback must actually inform model updates and organizational practices.

Transparency is also foundational. Architectures that allow for step-wise reasoning, like HRMs, or the use of confidence scores, can help build what Mittelstadt et al. (2016) call epistemic trust—the ability to know when and why to rely on a model. Ultimately, however, operational alignment hinges on people. It requires building critical AI literacy across the workforce, training employees to be mindful and alert to risks, not just effective prompt engineers. When governance tools, such as approval workflows and monitoring dashboards, are built directly into the enterprise platforms people already use, governance becomes less of an afterthought. It becomes an intrinsic part of the technological infrastructure itself (Ettinger, 2025).

4.3 Strategic Guardrails: Institutional Foresight and Ethical Anchoring

The third domain, Strategic Guardrails, is the institutional anchor for the other two. It consists of the high-level policies, structures, and values that ensure an organization's use of GenAI aligns with its mission, its legal obligations, and its societal role. In this view, ethical governance is not a checklist; it is an active, structural commitment. It requires dedicated bodies, such as Ethics Review Boards or AI Use Councils, to translate abstract principles into concrete policy and to adjudicate the inevitable edge cases (Crawford, 2021).

Because GenAI can be a powerful engine for amplifying systemic bias (Ferrara, 2023; Chen et al., 2023), these guardrails must include sophisticated tools for proactively finding and fixing it, from lifecycle assessments to BEATS-style audits (Lam et al., 2025). This oversight cannot be purely reactive. It must anticipate regulatory changes and build in the capacity for robust auditing and red-teaming from the start (Veale & Edwards, 2018). Key international standards, such as the EU AI Act, the NIST AI Risk Management Framework, and ISO/IEC 42001, provide a baseline for this type of risk-tiered governance. However, governance is not just an internal matter; it is also a reputational concern. Building public trust requires proactive engagement with customers, regulators, and civil society (Mittelstadt et al., 2016).

The most forward-looking function of this domain is what we call trajectory alignment. It is the capacity to govern not just the AI organizations possess today, but to anticipate how it might evolve (Jasanoff, 2016). With systems like ASIArch (Liu et al., 2025) on the horizon, which can recursively improve themselves, governance must expand to consider the trajectory of AI development itself. This is a critical challenge with profound implications for who controls the future of technology (Bostrom, 2014).

4.4 Reflexive Design and Domain Interdependence

These three domains are not a sequence or a checklist; they are a constant feedback loop. Feedback from Operational Alignment must inform the protocols of Model Stewardship. The vision set by Strategic Guardrails must shape both. This framework is designed to be reflexive—to learn and evolve in response to new technologies and new challenges (Beck, 1992). This is what it means to treat governance as a true organizational capability: dynamic, recursive, and deeply context aware.

5. The GenAI Governance Maturity Model (GAI-GMM)

From our analysis of existing IT maturity models and AI governance literature (detailed in Section 2.1), we developed the Generative AI Governance Maturity Model (GAI-GMM). This model outlines a five-stage journey, progressing from reactive, ad hoc usage to a state of proactive and resilient oversight. This model synthesizes the common pathways to maturity identified in our review and integrates them with our three-domain framework. We see maturity not as a final destination but as a process of building deeper capabilities across Model Stewardship, Operational Alignment, and Strategic Guardrails. The core idea is that as an organization matures, it moves from simply reacting to GenAI to intentionally shaping its trajectory.

The GAI-GMM's five stages of maturity—Ad Hoc / Experimental, Awareness / Initiation, Defined / Developing, Managed / Integrated, and Optimized / Reflexive—represent a progression of increasing sophistication in

governing GenAI. These stages are characterized by evolving capabilities in three key domains: Model Stewardship, Operational Alignment, and Strategic Guardrails. The model also maps four core governance capabilities—Intentionality, Integrity, Evolvability, and Reflexivity—across these domains and maturity levels to illustrate how each capability deepens and becomes more integrated as an organization matures. For example, Intentionality in the "Ad Hoc" stage is absent. In contrast, in the "Optimized" stage, it manifests through purposeful model selection and fine-tuning policies aimed at specific human-AI collaboration goals. Similarly, Integrity progresses from a reactive approach to bias detection to a proactive, lifecycle-spanning one. Evolvability and Reflexivity likewise mature from being non-existent to becoming a core institutional function. This progression from reactive to proactive governance is a central tenet of the model.

5.1.1 Stage 1: Ad hoc / experimental

This is the "Wild West" stage. Experimentation with GenAI is scattered and uncoordinated, happening in isolated pockets across the organization. There is no formal governance in place; teams utilize tools opportunistically, often for low-stakes tasks, with minimal documentation and review. There is no central model inventory, and most employees are likely unaware of any acceptable use policies. While this environment can spark initial creativity, it is rife with hidden risks —such as data leaks and reputational damage —that can escalate quickly.

5.1.2 Stage 2: Awareness / initiation

Here, the organization begins to realize that GenAI carries significant strategic and ethical weight. The conversation around governance typically starts here, often sparked by a minor incident or mounting regulatory pressure. The responsible team writes draft policies and forms informal working groups, but efforts remain fragmented. We may notice the beginnings of a model registry, though it is likely incomplete, and any user training is optional. The organization is starting to "sense" the governance challenge (Teece, 2007) but has not yet fully mobilized a unified response; discussions are often siloed in legal, IT, or ethics departments.

5.1.3 Stage 3: Defined / developing

At this stage, chaos begins to give way to order. The organization formalizes and communicates its core governance policies, with oversight mechanisms starting to stretch across the three domains. Formal approval workflows for new models are established, and the first human-in-the-loop (HITL) protocols are implemented for riskier applications. It is at this stage that the use of practical tools, such as bias audits or initial monitoring dashboards, becomes apparent. The organization is now "seizing" the opportunity to implement formal governance, though these new rules are often applied unevenly, focusing on a few high-risk use cases rather than the entire enterprise.

5.1.4 Stage 4: Managed / integrated

The fundamental transformation happens here. Governance is no longer just a policy document; the organization has successfully embedded it deeply in its work processes. Processes are not static but are regularly updated based on feedback and organizational learning, with end-to-end lifecycle management becoming the norm. The organization has adopted mandatory AI literacy training for employees and consistently uses enterprise platforms that incorporate built-in governance. At this point, the organization is "transforming" its own capabilities, making governance part of its operational DNA and moving beyond mere policy to a living, evolving system. It ceases to be a bureaucratic hurdle and instead becomes a source of competitive advantage and resilience.

5.1.5 Stage 5: Optimized / reflexive

It is the highest state of maturity, where governance is not just adaptive but predictive, deeply embedded in the organization's culture and strategy. It is capable of dynamically adjusting to new AI capabilities as they emerge and is institutionalized through advanced practices, such as governance sandboxes and red teaming. Here, we observe dedicated teams building systems with "governance-by-design" principles from the outset, and organizations often proactively engage in shaping public AI policy. In this stage, the organization has achieved a kind of epistemic coherence—a state in which its understanding of AI's capabilities, ethical principles, and governance mechanisms is constantly and dynamically aligned.

6. From Static to Reflexive: The Imperative of Dynamic Governance

Most corporate approaches to AI governance are built for a world that no longer exists, anchored in static compliance checklists and reactive audits. This old playbook is fundamentally broken in the face of modern

GenAI, which is emergent, adaptive, and, in some cases, capable of redesigning itself. The only way forward is a decisive shift from static control to a dynamic, reflexive capability. We define reflexive governance as the organization's capacity to constantly monitor, assess, and adapt its practices in an iterative loop (Beck, 1992). This approach reflects the logic of dynamic capabilities in action—the ability to feel emerging risks, seize opportunities, and transform the organization to institutionalize that learning (Teece, 2007). This adaptive posture requires building four interrelated organizational pillars.

First is leadership adaptability. Reflexive governance requires leaders to develop sophisticated AI fluency and treat governance as a central strategic issue. This involves empowering cross-functional councils and engaging in scenario-based planning to anticipate emergent capabilities, such as the self-evolution of ASI-Arch (Liu et al., 2025). This foresight is essential for managing the ultimate limits of human oversight (Bostrom, 2014).

Second, clear decision rights and accountability structures are required. Static models create bottlenecks; a reflexive approach strategically distributes decision rights to those closest to the technology while maintaining clear lines of accountability. Highly strategic decisions, particularly those involving autonomous reasoning systems, must rightly require formal board-level review (Kroll et al., 2017).

The third pillar is the creation of feedback-driven learning systems. Data fuels a reflexive system. This demands robust, real-time feedback loops that capture continuous monitoring data on model output drift or bias, systematic reports from employees, and proactive tracking of regulatory changes. This ensures governance is grounded in real-world impacts and allows the organization to absorb new information and adapt constantly.

Finally, proper cultural transformation is essential. Reflexive governance depends on a culture where employees are actively encouraged to question AI outputs and report problems without fear of reprisal. This marks a profound shift from a mindset of simply "adopting AI" to one of "stewarding AI responsibly." Ethical principles must become lived norms, reinforced through training and incentives that reward ethical foresight and responsible action.

This is not just a theoretical ideal; we are already seeing this kind of reflexivity in practice. The AstraZeneca case is a powerful example, and leading public institutions are pioneering tools such as AI use registries, governance sandboxes for high-risk experimentation, and third-party red teaming to test for vulnerabilities aggressively. This shift reframes AI governance as a dynamic institutional capability, one that provides a decisive advantage by enabling responsible innovation while safeguarding against the profound risks of this new technological era.

7. Implications for Practice and Policy

While every organization's journey with GenAI will be different, the need for strategic and ethical governance is universal. Based on our framework, we present a set of actionable recommendations for three key groups that are pivotal in shaping the future of this technology: C-suite leaders, policymakers, and internal ethics and governance committees.

7.1 For C-Suite Leaders

The ultimate responsibility for an organization's direction and risk posture lies with its C-suite, making their proactive engagement paramount for success. First, leaders must elevate GenAI governance to a core strategic priority. Second, leaders must commit to building their own fluency in GenAI's evolving capabilities and risks. Third, they must institutionalize cross-functional governance councils to break down the silos that so often hinder effective oversight.

7.2 For Policymakers and Regulators

Policymakers and regulators play a crucial role in shaping the environment for responsible AI innovation and public accountability. Their primary task should be to shift from rigid, prescriptive rules to principle-based frameworks. To foster innovation safely, they should develop regulatory sandboxes and risk-tiered standards. Sandboxes enable high-stakes experimentation in a controlled environment. At the same time, risk-tiered approaches apply the strictest oversight to high-impact domains, such as healthcare and finance, while allowing greater flexibility in other areas.

Given that AI is a global technology, it is essential to foster global regulatory interoperability by harmonizing standards across jurisdictions. Furthermore, regulators should mandate transparency in model provenance and deployment to build public trust and provide mechanisms for accountability. Finally, they should incentivize AI safety research and independent auditing by funding independent research and supporting a market for third-party audits and red-teaming.

7.3 For Internal AI Ethics Boards and Governance Committees

These internal bodies are the conscience and operational engine of AI governance, translating principles into practice. To be effective, they must first define a clear mandate and reporting structure that gives them genuine authority beyond a purely advisory role. Second, they must expand their expertise beyond purely technical domains. Their core function is to operationalize ethics through concrete policy and workflow, translating abstract values like fairness and transparency into enforceable model approval protocols and review criteria. Finally, these committees must build future readiness into their reviews.

8. Conclusion and Contribution

This paper presents a structured response to the governance challenges instigated by the arrival of GenAI. We proposed a conceptual framework built on three interlocking domains—Model Stewardship, Operational Alignment, and Strategic Guardrails—to manage the technical, operational, and ethical layers of the problem. To make this practical, we also introduced the GenAI Governance Maturity Model (GAI-GMM), a tool to help organizations benchmark their readiness and chart a path toward more resilient oversight.

Our theoretical contribution is to reconceptualize AI governance not as a regulatory burden, but as a dynamic organizational capability. This view, grounded in reflexivity and strategic foresight, sees governance as a source of adaptive advantage. For practice, this research offers clear guidance for executives, policymakers, and ethics boards, all of whom play a role.

9. Limitations and Avenues for Future Research

This paper presents a comprehensive conceptual framework for governing generative AI at scale. A primary limitation of this work is the absence of new empirical validation. While informed by illustrative cases and current developments discussed in the literature, the framework itself has not been subjected to primary data collection or rigorous empirical testing.

Furthermore, a significant proportion of the cutting-edge technological developments cited (e.g., ASI-Arch, HRMs, certain 2025 papers) are currently available as arXiv preprints and have not yet undergone full peer review. A reflexive critique of the institutionalist lens, while partially integrated, also warrants deeper exploration in future work, particularly concerning its potential to overlook non-institutional forms of governance or to perpetuate existing power structures (Eubanks, 2018; Crawford, 2021).

Ethics Declaration: This research did not require ethical clearance, as it did not involve human participants, animal subjects, or sensitive personal data.

AI Declaration: Artificial intelligence tools were used to support the research process, including assistance with the literature review, language editing, and stylistic refinement of the manuscript.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52161. https://doi.org/10.1109/ACCESS.2018.2870052

Aguilera, R. V., & Jackson, G. (2003). The cross-national diversity of corporate governance: Dimensions and determinants. Academy of Management Review, 28(3), 447–465.

Beck, U. (1992). Risk society: Towards a new modernity. Sage Publications.

Bhattacharya, A., Liang, P., & Narayanan, A. (2024). Explanatory debiasing: Leveraging domain expertise in data generation. arXiv preprint arXiv:2501.01441. https://arxiv.org/abs/2501.01441

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., … & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://arxiv.org/abs/2108.07258

Bostrom, N. (2014). Superintelligence: Paths, dangers, strategies. Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, S., … & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

Cath, C. (2018). Governing artificial intelligence: Ethical, legal and technical opportunities and challenges. Philosophical Transactions of the Royal Society A, 376(2133), 20180080.

Chen, F., Wang, X., Luo, J., & Yu, S. (2023). Unmasking bias in AI: A systematic review of bias detection and mitigation strategies in electronic health record–based models. arXiv preprint arXiv:2310.19917. https://arxiv.org/abs/2310.19917

Cohen, W. M., & Levinthal, D. A. (1990). Absorptive capacity: A new perspective on learning and innovation. Administrative Science Quarterly, 35(1), 128–152.

Crawford, K. (2021). Atlas of AI: Power, politics, and the planetary costs of artificial intelligence. Yale University Press.

Davenport, T. H., Guha, A., Grewal, D., & Bressgott, S. (2020). How artificial intelligence will change the future of marketing. Journal of the Academy of Marketing Science, 48(4), 1018–1043.

Eisenhardt, K. M., & Martin, J. A. (2000). Dynamic capabilities: What are they?. Strategic Management Journal, 21(10-11), 1105-1121.

Ettinger, R. (2025). Enterprise architecture as a dynamic capability for AI governance. arXiv preprint arXiv:2505.06326. https://arxiv.org/abs/2505.06326

EU AI Act. (2024). Official Journal of the European Union.

Eubanks, V. (2018). Automating inequality: How high-tech tools profile, police, and punish the poor. St. Martin's Press.

Ferrara, E. (2023). Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. arXiv preprint arXiv:2304.07683. https://arxiv.org/abs/2304.07683

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., … & Vayena, E. (2019). AI4People—Ethical guidelines for trustworthy AI: A European initiative. Minds and Machines, 29(4), 689–707.

Ganguli, D., Lohn, A., Chen, J., Dougal, S., & Shieber, S. (2022). Predictability and unreliability of large generative models during fine-tuning. arXiv preprint arXiv:2209.00667. https://arxiv.org/abs/2209.00667

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Metayer, D. L., & Crawford, K. (2018). Datasheets for datasets. arXiv preprint arXiv:1803.09010. https://arxiv.org/abs/1803.09010

Gibney, E. (2024). What the EU's tough AI law means for research and ChatGPT. Nature. https://doi.org/10.1038/d41586-024-00465-z

Giddens, A. (1990). The consequences of modernity. Stanford University Press.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudík, M., & Wallach, H. (2019). Improving fairness in machine learning systems: What do practitioners need from AI ethics research?. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 327-333.

ISO/IEC 42001:2023. (2023). Information technology—Artificial intelligence—Management system. ISO.

Jasanoff, S. (2016). The ethics of invention: Science and the responsible governance of innovation. Philosophy & Technology, 29(3), 295–298. https://doi.org/10.1007/s13347-016-0247-9

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705. [suspicious link removed]

Lam, C., et al. (2025). BEATS: A Risk-Based Assurance Audit Framework for Trustworthy Generative AI Systems. arXiv preprint arXiv:2507.11755. https://arxiv.org/abs/2507.11755

Liu, Y., Deng, W., Yu, X., & Pan, Z. (2025). ASI-Arch: Autonomous architecture generation and evaluation via large-scale self-optimization. arXiv preprint arXiv:2507.18074. https://arxiv.org/abs/2507.18074

MIT Sloan & Boston Consulting Group. (2025, August 22). Why 95% of enterprise AI fails—and what the 5% are doing right. Forbes.

Mitchell, M., Wu, S., et al. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3287040.3287044

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Science and Engineering Ethics, 22(5), 1187–1208. https://doi.org/10.1007/s13347-016-0284-6

Mökander, J., Floridi, L., Mittelstadt, B., & Schuett, J. (2024). Operationalizing AI governance in the biopharmaceutical sector: The AstraZeneca case. arXiv preprint arXiv:2407.05339. https://arxiv.org/abs/2407.05339

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2021). From what to how: An initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and Engineering Ethics, 27(1), 1–31. https://doi.org/10.1007/s11948-020-00277-x

NIST. (2023). AI Risk Management Framework 1.0. National Institute of Standards and Technology.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Powell, W. W., & DiMaggio, P. J. (Eds.). (1991). The new institutionalism in organizational analysis. University of Chicago Press.

Reuel, A., & Undheim, T. A. (2024). Generative AI needs adaptive governance. arXiv preprint arXiv:2406.04554. https://arxiv.org/abs/2406.04554

Scott, W. R. (1995). Institutions and organizations. Sage Publications.

Teece, D. J. (2007). Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. Strategic Management Journal, 28(13), 1319–1350. https://doi.org/10.1002/smj.640

Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting: An examination of the psychological situation and defence mechanisms of a work group. Human Relations, 4(1), 3–38. https://doi.org/10.1177/001872675100400101

Veale, M., & Edwards, L. (2018). Clarity, control and contestability in algorithmic regulation. Computer Law & Security Review, 34(4), 795-805. https://doi.org/10.1016/j.clsr.2018.03.003

Wang, G., Li, J., Sun, Y., & Chen, X. (2025). Hierarchical Reasoning Model. arXiv preprint arXiv:2506.21734. https://arxiv.org/abs/2506.21734