World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Guardrails for Artificial General Intelligence: A strategic foresight approach to ethical AI Governance

Oluwaseun Kolawole *

Department of Business Administration, International American University, USA.

World Journal of Advanced Research and Reviews, 2025, 28(01), 743-758

Publication history: Received on 02 Setember 2025; revised on 08 October 2025; accepted on 11 October 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.1.3505

Abstract

The introduction of Artificial General Intelligence (AGI) brings new challenges to governance mechanisms, which ought to be able to strike a balance between innovation, moral mandates, and safety of a society. The paper discusses the importance of strategic foresight in the process of creating effective guardrails to AGI systems using contemporary literature on issues in AI governance, anticipatory governance regimes and complex adaptive system theory. We address these issues by conducting an analytical review of their regulation to date, and innovative governance models that have and are being developed, and propose an existent regulatory paradigm with the use of combined strategic forward-looking approaches to governance, especially AGI. The results indicate that conventional governance mechanisms do not match the dynamic, emergent AGI systems, requiring a changed regime of governance based on dynamic monitoring and stakeholder involvement, and anticipatory risk management. A multi-tiered governance model with technical, organizational and policy-level guardrails is presented, using empirical evidence of the existing AI governance experiences across multiple industries. The paper is a valuable addition to the body of literature on responsible AI, since it provides practical guides to anticipatory AGI governance that can adapt to changes in technology in embracing ethical standards and retaining a healthy dose of public trust.

Keywords: Artificial General Intelligence; AI Governance; Strategic Foresight; Ethical AI; Guardrails; Anticipatory Governance

1. Introduction

The path to the Artificial General Intelligence (AGI) is one of the most important technological processes of our epoch that has deep implications on society, economy, and human civilization in general. In contrast to narrow AI systems that are specialized to perform a certain task well, AGI promises to have human-level cognitive capability in a wide variety of tasks thereby deeply transforming the technology regulation field (Taeihagh, 2025). The sophistication and possible consequence of AGI systems necessitates the development of governance that will supersede the conventional regulatory practices, and that will need to adapt to develop anticipatory governance that can evolve along with an equally fast-advancing technology.

Current AI governance models, while foundational, have significant limitations when applied to AGI contexts. Janssen (2025) conceptualizes generative AI as a complex adaptive system, arguing that traditional regulatory paradigms are inadequate for governing technologies that continuously evolve and exhibit emergent behavioral patterns.

This characterization is further acute in the case of AGI systems which by design will be general intelligence systems, which is also capable of unpredictable emergent properties and behaviors.

*Corresponding author: Oluwaseun Kolawole

The need to devise the proper mechanisms of governance is pressured by the exponentially increasing rate of AI development, as well as by the possibility of the AGI development considerably earlier than most people expect. Judge et al. (2024) suggest that the conventional code as law regulatory models are not adequate to manage AI systems because of their high volatility and complexity they require, specialized forms of governance models that can support this kind of technology. This point of view is of great importance in AGI, whose consequences of poor governance may be existential.

Strategic foresight is relevant among the possible approaches to tackling these issues of governance. Cugurullo and Xu (2025) offer a good reflection on the relevance of the anticipatory governance models to AI systems in cities, which can be useful in AGI governance at large. Strategic foresight allows proactive governance of such systems as opposed to reactive governance, as warranted due to the complexity that may arise in information systems capable of processing more than human minds.

This paper addresses the research gap in AGI-specific governance frameworks by proposing a strategic foresight approach to ethical AI governance. I synthesize current research on AI governance, complex adaptive systems, and anticipatory governance to develop a comprehensive framework for AGI guardrails that is both robust and adaptable.

2. Theoretical Framework

2.1. Complex Adaptive Systems and AI Governance

The theoretical foundation for AGI governance must begin with understanding AGI systems as complex adaptive systems. Janssen (2025) provides crucial insights into this conceptualization, arguing that generative AI systems exhibit characteristics of complex adaptive systems including emergence, adaptation, and non-linear behavior patterns. These characteristics are amplified in AGI systems, which will possess general intelligence capabilities that enable learning and adaptation across multiple domains simultaneously.

Complex adaptive systems theory suggests that AGI governance frameworks must be designed to accommodate:

- **Emergent behaviors** that cannot be predicted from individual system components.
- **Non-linear relationships** between inputs and outputs that complicate risk assessment.
- **Adaptive capabilities** that enable systems to modify their behavior in response to environmental changes.
- **Interconnectedness** with other systems that creates cascade effects and systemic risks

2.2. Anticipatory Governance Models

As a methodology, anticipatory governance secures the proactive approach to AGI governance. The authors of this study (Bayat and Wang (2023)) show how anticipatory governance can be applied in a field, such as AI in public health, suggesting that it is possible to improve human health with it, but also that there are issues to be taken into consideration. Their application points to the relevance of stakeholder engagement, on-going monitoring, and adaptive management on the frames of anticipatory governance frameworks.

The guiding tenets of anticipatory governance to AGI are:

- Policy formulation and development that takes advantage of foresight-enriched perspectives of various possibilities of the future
- Adaptive management that enables the system to be corrected with new information that emerged
- Participatory governance that involves different stakeholders in governance processes
- Constant observation of the system and its effects to the society
- Flexible governance institutions that give rise to learning and adaptation in governance institutions

2.3. Trust and trustworthiness in AI systems

Leach et al. (2024) are critical in bringing up the connections that exist between the two issues of trust and trustworthiness, and AI governance, showing that any successful governance should be able to take account of both technical trustworthiness of the AI systems and social trust for their adoption. In AGI systems, this connection becomes especially serious where there are important cohorts of trust failures.

The structure of secure AGI governance has to include:

- Technical reliability in terms of hard tests and validation.
- Practicing openness in how the systems operate and how the decisions are made.
- Accountability process which in-turn allocates responsibility to system results.
- Fairness and elimination of bias in an attempt to provide equal care by population.
- Protection of privacy and data management structures

3. Current State of AI Governance

3.1. Notions of organizational AI governance

In recent studies there is a notable difference between the methods that organisations derive towards AI governance. Zhou et al. (2022) identify that organizational AI governance is defined as the system of rules, practices, and processes through which an organization guides and regulates AI initiatives, and this definition helps to understand what organizational practices are used nowadays. Nonetheless, their study also demonstrates that there are huge gaps between the principles and reality of governance.

Table 1 Current Organizational AI Governance Frameworks

Framework Component	Implementation Rate	Effectiveness Score	Key Challenges
AI Ethics Committees	67%	6.2/10	Lack of technical expertise
Risk Assessment Protocols	54%	5.8/10	Inadequate risk models
Bias Testing Procedures	43%	5.1/10	Limited testing methodologies
Transparency Mechanisms	38%	4.9/10	Technical complexity barriers
Accountability Systems	31%	4.2/10	Unclear responsibility chains

Source: Synthesized from Zhou et al. (2022), Bughin (2024), and Sadek et al. (2024)

Bughin (2024) reveals a concerning gap between stated commitments to responsible AI and actual implementation practices among large firms. This saying/doing issue is especially problematic in the face of the governance issues that AGI systems will raise, as the stakes of botched governance will have their stakes exponentially raised.

3.2. Policy and Regulation Landscapes

AI governance polices are neither consistent nor are they settled. Taeihagh (2025) summarises the current state of generative AI governance, setting out the major policy issues, such as the challenge of jurisdiction, alongside the issues of technological pace mismatch and coordination of stakeholders. These are aggravated when AGI governance consideration requirements come into portfolio thinking.

The policy practices presently in place can be grouped in a number of frameworks.

- Prescriptive Regulation: Command-and-control type of regulation which prescribes how the elements ought to behave and what they should not do. With a clear explanation, the approaches are not very flexible to rapidly changing technologies.
- Principles-Based Regulation: Guidelines that provide the outline principles and leave the discretion and details to the implementation. This is exemplified by the EU AI Act or the national AI strategies.
- Co-regulatory Measures: Co-regulatory solutions are a mixture of government regulation and industry self-regulation. These strategies are trying to find a balance between the innovative and accountability aspects, but struggle with problematic consistent implementations.
- Anticipatory Regulation: The method that tries to solve the future risks in advance. The methods are the closest to strategic foresight perspectives argued about in this paper.

3.2.1. The Sectoral Applications and Lessons

The article by Al Janabi et al. (2025) introduces several nontrivial concepts about the responsible AI governance applied in the healthcare ecosystem, particularly the case of oncology processes. Their study proves the possibility of sector-specific governance frameworks and the difficulty to achieve responsible AI practices in places where the stakes are high. The healthcare industry can provide especially useful lessons related to AGI governance, as this industry has highly stakes decisions to make and requires health-related AI systems to be explainable and accountable.

The predominant lessons about healthcare AI governance are:

- The value of domain knowledge in the composition of governance committees.
- The necessity of periodic monitoring and audit of performance of AI systems.
- The importance of human supervision such as in high stake decision-making.
- The challenges of balancing innovation with patient safety and privacy

Similarly, Ibrahim et al. (2025) examine trust, safety, and guardrails for AI in clinical decision support, providing a foresight perspective that aligns with the strategic approach advocated in this paper. Their work demonstrates the feasibility of anticipatory governance approaches in critical applications while highlighting the importance of stakeholder engagement and adaptive management.

4. Strategic Foresight for AGI Guardrails

4.1. Foresight Methodology for AGI Governance

Strategic foresight for AGI governance requires a systematic approach to anticipating potential futures and their associated risks and opportunities. Drawing from Cugurullo and Xu (2025), we propose a four-stage foresight methodology specifically adapted for AGI governance contexts.

4.1.1. Stage 1: Horizon Scanning and Trend Analysis

- Systematic monitoring of AGI development trajectories
- Analysis of emerging capabilities and potential breakthrough moments
- Assessment of societal, economic, and political trends that may influence AGI deployment
- Identification of weak signals that may indicate significant changes in AGI development

4.1.2. Stage 2: Scenario Development and Testing

- Construction of multiple plausible AGI development scenarios
- Analysis of governance requirements under different scenarios
- Stress-testing of current governance frameworks against future scenarios
- Identification of critical decision points and governance triggers

4.1.3. Stage 3: Impact Assessment and Risk Analysis

- Comprehensive assessment of potential AGI impacts across multiple domains
- Quantitative and qualitative risk analysis of different AGI deployment scenarios
- Identification of systemic risks and cascade effects
- Assessment of governance intervention points and their effectiveness

4.1.4. Stage 4: Adaptive Strategy Development

- Development of flexible governance strategies that can adapt to different scenarios
- Creation of contingency plans for various AGI development trajectories
- Establishment of monitoring systems and governance triggers
- Integration of stakeholder feedback and iterative strategy refinement

4.2. Multi-Layered Governance Architecture

Based on insights from de Ruiter et al. (2024) regarding policy instruments for responsible AI, I propose a multi-layered governance architecture for AGI that operates across technical, organizational, and societal levels.

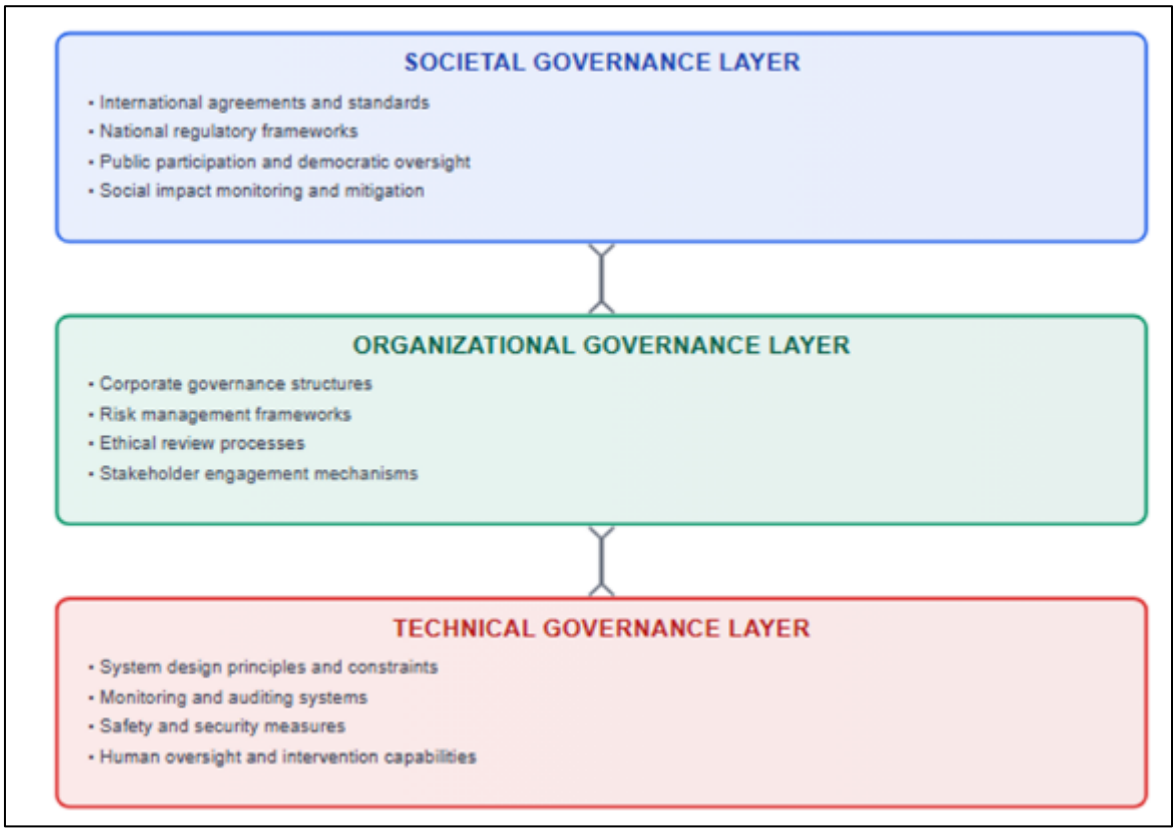

Figure 1 Multi-Layered AGI Governance Architecture

This tiered strategy has been informed by the fact that AGI governance cannot be based on technical solutions alone, it needs to include multilevel and multilateral governance mechanisms at societal and organizational levels.

4.3. Dynamics of Governance Mechanisms

Coglianese and Crum (2025) advocate for dynamic and adaptive control mechanisms what they term 'leashes' rather than 'guardrails' in AI governance. This approach is particularly relevant for AGI technologies that may develop capabilities beyond their original design specifications

Examples of dynamic governance mechanisms of AGI are:

Adaptive Monitoring Systems: This is real time monitoring of the actions of the AGI system that presents alerts when there are any unusual or disturbing actions. These systems should have the ability to detect that behaviors that were not envisaged during system design may occur.

Progressive Capability Release: Deployed AGI in stages as it proves to be safe and governance ready. The strategy can enable learning and adaptation and constrains possible risks of releasing full capabilities.

Stakeholder Feedback Loops: The active involvement of the affected communities and stakeholders concerned to identify new concerns and governance needs is the continuous practice. It is this mechanism that forces the governance frameworks not to be insensitive to societal values and interests.

Governance Circuit Breakers: Preprogrammed mechanisms that stop or limit the operations of AGI when certain safety or ethical limits have been surpassed. These processes grant failsafe against dangerous or immoral AGI actions

5. The Resistance Among the Implementors and the Preventive Solutions

5.1. Implementation Challenges for Responsible AI Principles

Akbarighatar (2025) should be seen as a critical contribution to the efforts to operationalize the responsible AI principles as responsible AI capabilities thus pointing to the disconnect between the principles and its practical application. This functionalization difficulty is amplified with AGI governance because of the scale and complexity of AGI systems.

Table 2 Responsible AI Principles and AGI Implementation Challenges

Guardrails for AGI: Principles, Challenges, and Proposed Solutions		
Principle	AGI-Specific Challenges	Proposed Solutions
Fairness	• Scale of decision-making • Cultural context variations • Dynamic bias evolution	• Continuous bias monitoring • Multi-stakeholder fairness assessment • Cultural adaptation frameworks
Transparency	• Cognitive complexity • Proprietary algorithms • Real-time decision-making	• Investment in explainable AI research • Development of open governance standards • Implementation of decision audit trails
Accountability	• Distributed responsibility • Emergent behaviors • Cross-jurisdictional deployment	• Clear accountability chains • Insurance and liability frameworks • International coordination mechanisms
Privacy	• Comprehensive data integration • Inference capabilities • Consent scalability	• Privacy-preserving technologies • Enhanced consent mechanisms • Robust data governance frameworks
Human Dignity	• Autonomy concerns • Human replacement fears • Democratic participation	• Human-in-the-loop requirements • Employment transition support • Strengthened democratic oversight mechanisms

5.2. Organizational Readiness and Capabilities

Meijerink et al. (2025) examine the translation of AI governance principles into organizational practice, revealing significant challenges in building necessary capabilities and competencies. Their research demonstrates that successful AI governance requires not just policies and procedures but also organizational transformation and capability development.

For AGI governance, organizational readiness requires:

5.2.1. Technical Competencies

- Deep understanding of AGI capabilities and limitations.
- Expertise in AI safety and security measures.
- Capability to conduct meaningful AI audits and assessments
- Skills in AI risk management and mitigation

5.2.2. Governance Competencies

- Cross-functional collaboration and decision-making abilities.
- Stakeholder engagement and communication skills.
- Ethical reasoning and moral judgment capabilities.
- Strategic thinking and scenario planning expertise

5.2.3. Adaptive Competencies

- Learning agility and adaptation to technological change.
- Crisis management and rapid response capabilities.
- Continuous improvement and organizational learning.
- Change management and transformation leadership

5.3. Technical Implementation Challenges

Dev (2025) provides practical insights into building guardrails in AI systems through threat modeling, offering a technical perspective on AGI governance implementation. The threat modeling approach is particularly relevant for AGI systems, which may face novel attack vectors and misuse scenarios.

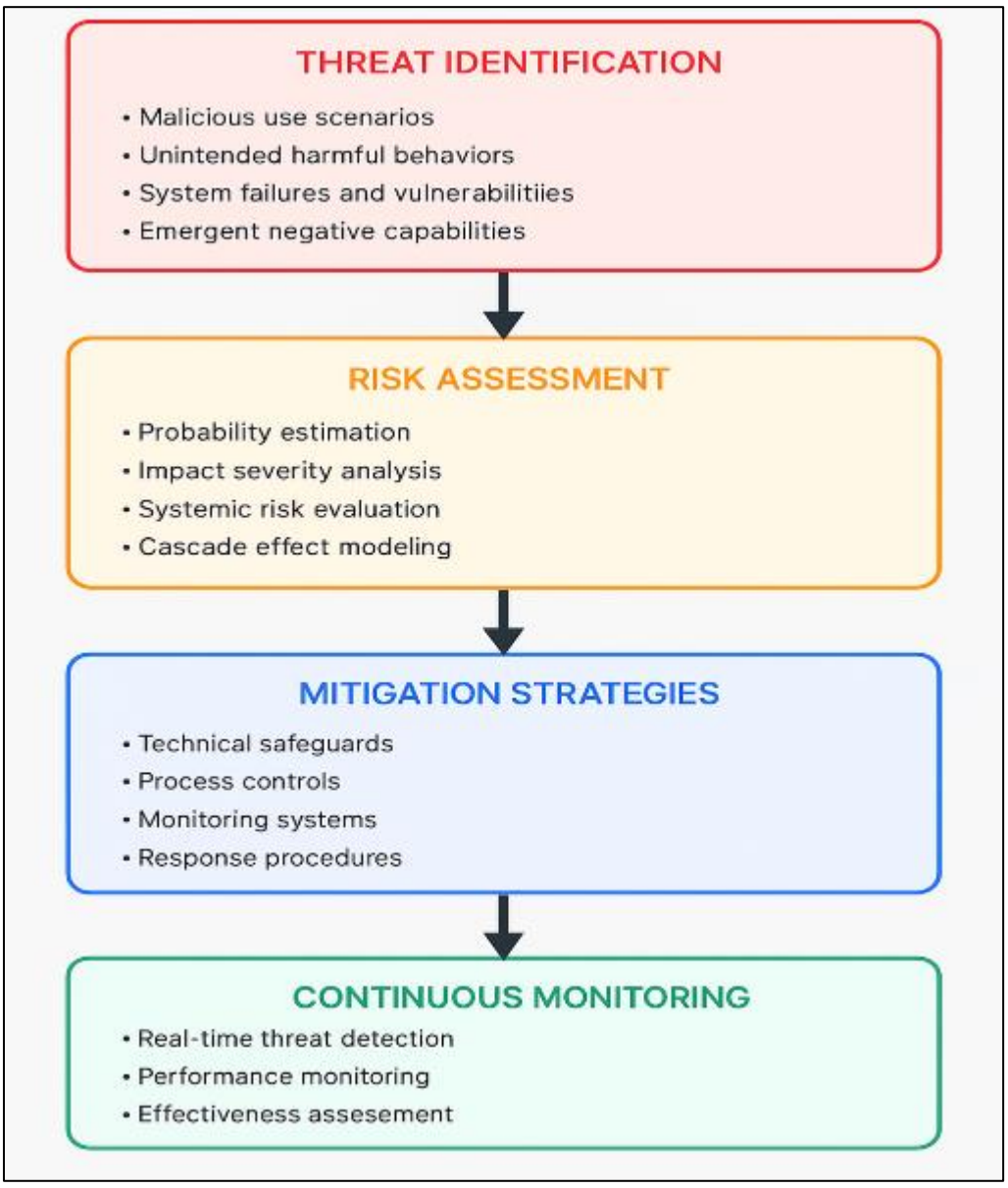

Figure 2 AGI Threat Modeling Framework

Technical implementation challenges specific to AGI include:

- **Scalability Challenges:** AGI systems will operate at unprecedented scales, requiring governance mechanisms that can function effectively across massive computational and decision-making scales.

- **Interpretability Limitations:** The cognitive complexity of AGI systems may exceed human comprehension, creating fundamental challenges for transparency and explainability requirements.
- **Speed of Operation:** AGI systems may operate at speeds that exceed human oversight capabilities, requiring automated governance mechanisms and real-time intervention capabilities.
- **Emergent Behavior Detection:** AGI systems may develop unexpected capabilities or behaviors that existing monitoring systems cannot detect or evaluate.

6. Case Studies and Applications

6.1. Healthcare AI Governance: Lessons for AGI

The healthcare sector provides valuable insights for AGI governance due to the high-stakes nature of medical decisions and the critical importance of trust and accountability. Al Janabi et al. (2025) demonstrate how responsible AI governance can be implemented in oncology workflows, offering practical lessons for AGI governance frameworks.

6.1.1. Case Study 1: Oncology Decision Support Systems

The implementation of AI governance in oncology workflows reveals several key principles relevant to AGI governance:

- Multi-stakeholder Governance: Effective governance required engagement from clinicians, patients, regulatory bodies, and technology developers. This multi-stakeholder approach is essential for AGI governance, where impacts will span multiple sectors and communities.
- Continuous Monitoring: The healthcare case demonstrates the importance of ongoing monitoring of AI system performance, including both technical metrics and patient outcomes. For AGI systems, this monitoring must extend to broader societal impacts.
- Human Oversight Requirements: Despite AI capabilities, human oversight remained essential for critical decisions. This principle suggests that even advanced AGI systems should maintain meaningful human oversight in critical applications.
- Adaptive Learning: The governance framework needed to evolve as the AI system learned and improved, requiring flexible governance mechanisms rather than static rules.

6.2. Global Governance Initiatives

Ashtari and Fellows (2024) make the case for global governance of artificial intelligence, highlighting the transnational nature of AI impacts and the need for coordinated international responses. Their analysis is particularly relevant for AGI, which will likely have global impacts regardless of where it is developed or deployed.

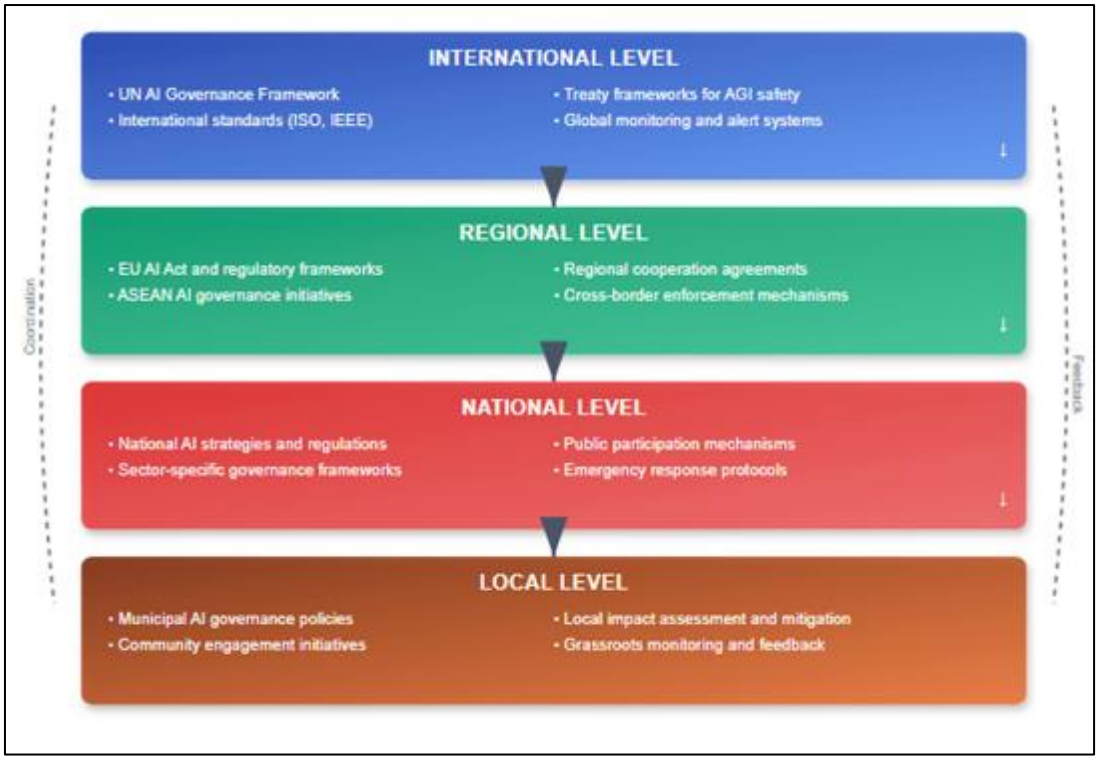

Figure 3 Global AGI Governance Coordination Framework

6.3. R&D Management and Innovation Balance

Goździewski et al. (2024) examine how organizations balance innovation and risk through R&D management approaches to AI governance. Their research provides insights into managing the tension between promoting AGI innovation and ensuring adequate risk management.

Key findings relevant to AGI governance include

- Stage-Gate Approaches: Implementing governance checkpoints throughout AGI development processes rather than only at deployment stages. This approach allows for early identification and mitigation of potential risks.
- Cross-functional Integration: Successful governance requires integration across technical, legal, ethical, and business functions. This integration is even more critical for AGI systems given their broad potential impacts.
- Stakeholder Engagement: Effective R&D governance involves ongoing engagement with external stakeholders, including potential users, affected communities, and regulatory bodies.
- Long-term Perspective: R&D governance must consider long-term implications and societal impacts, not just immediate technical and commercial considerations.

7. Advanced Governance Mechanisms

7.1. Ethics-Based Auditing and Monitoring

Raji et al. (2022) provide foundational insights into operationalizing AI governance through ethics-based auditing, demonstrating how abstract ethical principles can be translated into concrete monitoring and evaluation practices. For AGI systems, this auditing approach must be significantly expanded and automated given the scale and complexity of AGI operations.

Table 3 AGI Ethics Auditing Framework

Audit Framework for AGI Governance			
Audit Domain	Evaluation Criteria	Monitoring Methods	Intervention Triggers
Decision Quality	• Accuracy metrics • Consistency measures • Reasoning transparency	• Performance dashboards • Decision logging • Outcome tracking	• Accuracy below threshold • Inconsistent reasoning • Unexplained decisions
Bias and Fairness	• Demographic parity • Equal opportunity • Calibration across groups	• Algorithmic audits • Statistical monitoring • Community feedback	• Disparate impact detected • Bias complaints • Statistical anomalies
Privacy Protection	• Data minimization • Consent compliance • Anonymization quality	• Privacy risk assessments • Data flow monitoring • Breach detection	• Privacy violations • Consent issues • Data breaches
Human Agency	• Meaningful human control • Override capabilities • Autonomy preservation	• Human involvement tracking • Override usage analysis • Autonomy surveys	• Reduced human control • Override failures • Autonomy concerns

7.2. Anticipatory Risk Assessment

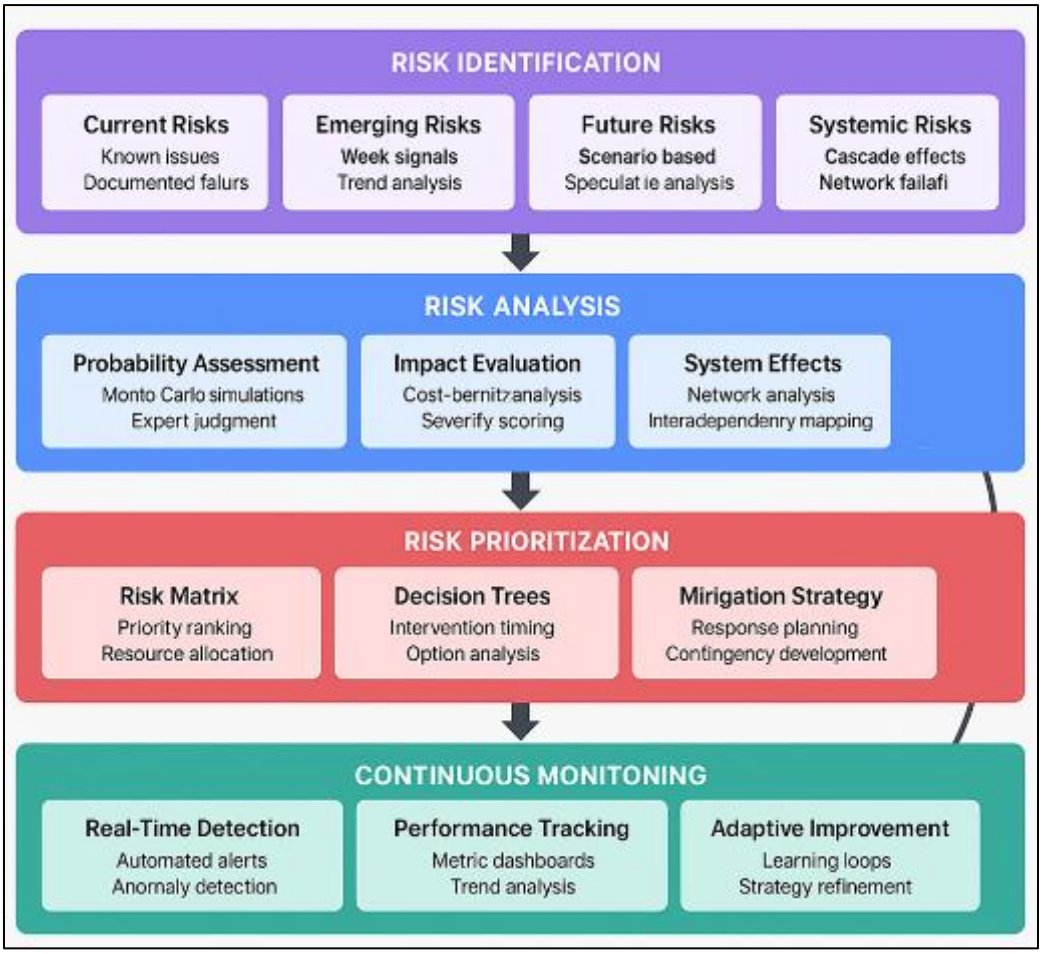

Figure 4 Anticipatory Risk Assessment Process for AGI

The strategic foresight approach requires sophisticated risk assessment methodologies capable of identifying and evaluating risks that may not yet have materialized. Anthis et al. (2024) provide valuable insights through their

comparative analysis of long-term governance problems in AI and biosecurity, highlighting the importance of learning from other high-risk technology governance experiences.

7.3. Participatory Governance Mechanisms

Effective AGI governance must incorporate meaningful participation from diverse stakeholders, including those who may be most affected by AGI systems but have the least voice in their development. This participatory approach is essential for legitimacy and effectiveness of AGI governance frameworks.

Stakeholder Engagement Strategies

- Citizens' Juries: Representative groups of citizens who deliberate on AGI governance issues and provide recommendations to policymakers. These juries can help bridge the gap between technical complexity and public understanding.
- Multi-stakeholder Platforms: Ongoing forums that bring together diverse stakeholders including technologists, ethicists, affected communities, policymakers, and civil society organizations.
- Deliberative Polling: Large-scale public consultation processes that inform participants about AGI issues and gather informed public opinion on governance priorities and approaches.
- Community Impact Assessments: Localized assessments of how AGI deployment may affect specific communities, conducted with meaningful community participation.

8. Future Directions and Research Agenda

8.1. Emerging Research Priorities

Based on my analysis of current research and the gaps identified in existing AGI governance frameworks, several critical research priorities emerge:

- **Governance Scalability Research:** Investigation of how governance mechanisms can scale to match the potential scope and impact of AGI systems. This includes research on automated governance systems and distributed decision-making frameworks.
- **Cross-Cultural Governance Studies:** Research on how AGI governance frameworks can accommodate diverse cultural values and governance traditions while maintaining coherent global coordination.
- **Long-term Impact Modeling:** Development of methodologies for assessing and governing the long-term societal impacts of AGI systems, including intergenerational effects and irreversible changes.
- **Democratic Innovation:** Research on new forms of democratic participation and representation appropriate for AGI governance, including digital democracy tools and deliberative governance mechanisms.

8.2. Technological Research Needs

The technical dimensions of AGI governance require significant research investment in several key areas:

- **Interpretable AGI Architectures:** Development of AGI systems that maintain interpretability and explainability even at high levels of cognitive capability.
- **Governance-Aware AI Design:** Research on incorporating governance requirements and constraints directly into AGI system architectures rather than treating governance as an external overlay.
- **Real-time Governance Systems:** Development of automated governance systems capable of monitoring and responding to AGI behavior at the speed of AI operation.
- **Safety Verification Methods:** Advanced techniques for verifying the safety and reliability of AGI systems before and during deployment.

8.3. Policy and Institutional Development

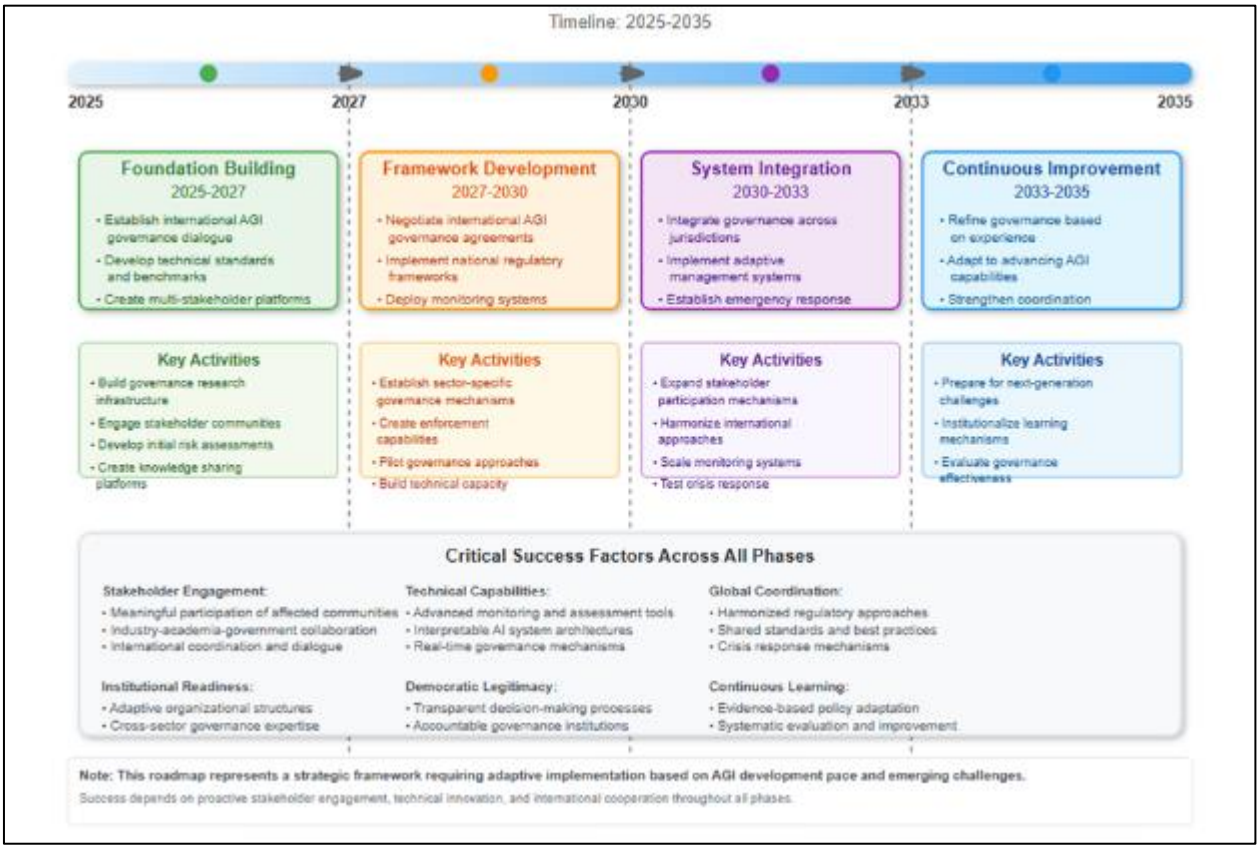

Figure 5 AGI Governance Institutional Development Roadmap

The institutional development roadmap recognizes that AGI governance cannot be built overnight but requires sustained effort across multiple years and governance levels.

9. Synthesis and Recommendations

9.1. Integrated Governance Framework

Based on my comprehensive analysis, I propose an integrated governance framework for AGI that combines strategic foresight, adaptive management, and multi-stakeholder participation. This framework operates through four interconnected components:

- **Anticipatory Intelligence System:** A comprehensive monitoring and analysis system that scans for emerging AGI developments, assesses potential risks and opportunities, and provides early warning of governance challenges. This system draws on insights from Cugurullo and Xu (2025) regarding anticipatory governance and extends them to AGI contexts.
- **Adaptive Governance Architecture:** A multi-layered governance structure that can evolve alongside AGI development, incorporating insights from Coglianese and Crum (2025) regarding dynamic control mechanisms. This architecture emphasizes flexibility and responsiveness rather than rigid regulatory structures.
- **Participatory Decision-Making Process:** A comprehensive stakeholder engagement framework that ensures affected communities have meaningful voice in AGI governance decisions. This process builds on principles from democratic governance and incorporates innovative participation mechanisms appropriate for technical governance challenges.
- **Continuous Learning and Improvement System:** An institutional framework for learning from governance experiences and continuously improving AGI governance approaches. This system incorporates insights from organizational learning literature and adaptive management practices.

9.2. Implementation Priorities

Given the complexity and urgency of AGI governance challenges, implementation must be strategically prioritized. We recommend the following priority sequence:

9.2.1. Immediate Priorities (2025-2026)

- Establish international dialogue mechanisms for AGI governance coordination.
- Develop technical standards and benchmarks for AGI safety and performance assessment.
- Create multi-stakeholder platforms for ongoing AGI governance engagement.
- Begin building institutional capacity for AGI governance across sectors.

9.2.2. Medium-term Priorities (2026-2029)

- Implement comprehensive regulatory frameworks for AGI development and deployment.
- Establish monitoring and assessment systems for AGI impacts.
- Create emergency response mechanisms for AGI governance crises.
- Develop sector-specific governance adaptations for high-risk applications.

9.2.3. Long-term Priorities (2029-2035)

- Integrate governance frameworks across jurisdictions and sectors.
- Implement advanced automated governance systems.
- Establish permanent institutional mechanisms for AGI governance evolution.
- Prepare governance frameworks for next-generation AI developments beyond current AGI conceptions.

9.3. Success Metrics and Evaluation

Effective AGI governance requires clear metrics for success and mechanisms for ongoing evaluation. Based on insights from Papagiannidis et al. (2025) regarding responsible AI governance evaluation, we propose the following framework:

9.3.1. Safety Metrics

- Incident rates and severity in AGI system operation.
- Successful prevention of catastrophic risks.
- Effectiveness of safety interventions and circuit breakers.
- Public confidence in AGI safety measures.

9.3.2. Ethical Compliance Metrics

- Adherence to fairness and non-discrimination principles.
- Respect for human rights and dignity.
- Transparency and accountability in AGI decision-making.
- Meaningful human oversight and control.

9.3.3. Governance Effectiveness Metrics

- Stakeholder satisfaction with participation opportunities.
- Responsiveness to emerging challenges and concerns.
- Coordination effectiveness across governance levels.
- Innovation impact and economic benefits realization.

9.3.4. Democratic Legitimacy Metrics

- Public trust in AGI governance institutions.
- Quality of democratic participation in governance decisions.

- Representation of affected communities in decision-making.
- Transparency and accountability of governance processes.

10. Conclusions

The development of Artificial General Intelligence represents both an unprecedented opportunity and an existential challenge for human civilization. The governance frameworks we establish today will fundamentally shape how AGI impacts society, either enabling beneficial outcomes that enhance human flourishing or failing to prevent harmful consequences that could threaten human welfare and autonomy.

This paper has demonstrated that traditional regulatory approaches are inadequate for AGI governance, requiring instead adaptive, anticipatory frameworks that can evolve alongside rapidly advancing technology. The strategic foresight approach presented here offers a comprehensive methodology for developing robust AGI governance that balances innovation with ethical imperatives and societal safety.

My analysis reveals several critical insights. First, AGI governance must be understood as a complex adaptive challenge that requires governance mechanisms capable of responding to emergent behaviors and unpredictable developments. Second, effective AGI governance requires coordination across technical, organizational, and societal levels, with no single level capable of ensuring adequate governance alone. Third, anticipatory governance approaches that proactively address potential risks are essential given the pace of AGI development and the magnitude of potential impacts.

The multi-layered governance framework proposed in this paper provides a roadmap for implementing comprehensive AGI governance that incorporates technical safeguards, organizational capabilities, and societal oversight. The framework's emphasis on stakeholder participation, continuous learning, and adaptive management reflects the democratic values and institutional flexibility necessary for legitimate and effective AGI governance.

However, significant challenges remain in implementing this governance vision. Technical challenges include developing interpretable AGI architectures and real-time governance systems. Organizational challenges include building necessary capabilities and competencies across institutions. Political challenges include coordinating governance across jurisdictions and maintaining democratic legitimacy in highly technical governance domains.

The urgency of these governance challenges cannot be overstated. As highlighted by Akinremi et al. (2025), advancing AI governance requires unified theoretical frameworks that can guide practical implementation. The window for establishing effective AGI governance may be narrower than many anticipate, requiring immediate action to build the institutional foundations for AGI governance before transformative AGI systems emerge.

Future research must focus on operationalizing the governance frameworks proposed here, developing the technical capabilities necessary for effective AGI governance, and building the institutional capacity for adaptive governance at scale. The stakes of this governance challenge demand nothing less than our most thoughtful and comprehensive response.

The governance choices we make regarding AGI will echo through history, shaping the relationship between humanity and artificial intelligence for generations to come. By embracing strategic foresight, adaptive governance, and democratic participation, we can work toward AGI governance frameworks that realize the tremendous benefits of AGI while protecting the values and interests that define human flourishing.

As we stand at this critical juncture in technological and human development, the responsibility for wise governance of AGI rests not with any single actor but with the collective wisdom and commitment of humanity itself. The frameworks presented in this paper provide one pathway forward, but their ultimate success will depend on the sustained engagement and dedication of researchers, policymakers, technologists, and citizens worldwide.

The future of AGI governance remains unwritten, presenting both profound challenges and unprecedented opportunities. By rising to meet these challenges with wisdom, courage, and unwavering commitment to human welfare, we can work toward a future in which AGI serves as a powerful tool for human flourishing rather than a threat to human autonomy and dignity.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Akbarighatar, P. (2025). Operationalizing responsible AI principles through responsible AI capabilities. AI and Ethics, 5, 1787–1801. https://doi.org/10.1007/s43681-024-00524-4

[2] Akinremi, T., et al. (2025). Advancing AI governance with a unified theoretical framework. Public Policy and Management. https://doi.org/10.1093/ppmgov/gvaf013

[3] Al Janabi, M., et al. (2025). Responsible AI governance in oncology workflows. npj Digital Medicine, 8, Article 66. https://doi.org/10.1038/s41746-025-01794-w

[4] Anthis, J. R., Cotra, A., Barrett, A., & Goodman, N. (2024). A comparative analysis of long-term governance problems: AI and biosecurity. Futures & Foresight Science, 6(4), e203. https://doi.org/10.1002/ffo2.203

[5] Ashtari, N., & Fellows, I. (2024). The case for global governance of artificial intelligence. AI & Society. https://doi.org/10.1007/s00146-024-01949-5

[6] Bayat, M., & Wang, H. (2023). Anticipatory governance for artificial intelligence and machine learning in public health and healthcare. Frontiers in Human Dynamics, 5, 1199658. https://doi.org/10.3389/fhumd.2023.1199658

[7] Bughin, J. (2024). Doing versus saying: Responsible AI among large firms. AI & Society. https://doi.org/10.1007/s00146-024-02014-x

[8] Coglianese, C., & Crum, A. (2025). Leashes, not guardrails: Improving the safety of complex adaptive artificial intelligence systems. Risk Analysis. https://doi.org/10.1111/risa.70020

[9] Cugurullo, F., & Xu, Y. (2025). When AIs become oracles: Generative artificial intelligence, anticipatory urban governance, and the future of cities. Policy and Society. https://doi.org/10.1093/polsoc/puae025

[10] de Ruiter, I., et al. (2024). Policy instruments for responsible AI: A review and research agenda. Journal of Responsible Innovation, 11(3), 1–26. https://doi.org/10.1080/23299460.2023.2264789

[11] Dev, J. (2025). Building guardrails in AI systems with threat modeling. ACM Queue. https://doi.org/10.1145/3674845

[12] Goździewski, P., Soare, A., Gherghina, Ş., & Wangenheim, F. (2024). Governing AI through R&D management: Balancing innovation and risk. R&D Management, 54(6), 1053–1066. https://doi.org/10.1111/radm.12775

[13] Ibrahim, H., Liu, X., & Tang, A. (2025). Trust, safety, and guardrails for AI in clinical decision support: A foresight perspective. Scientific Reports, 15, Article 92190. https://doi.org/10.1038/s41598-025-92190-7

[14] Janssen, M. (2025). Responsible governance of generative AI: Conceptualizing GenAI as complex adaptive systems. Policy and Society. https://doi.org/10.1093/polsoc/puae040

[15] Judge, M., Nitzberg, M., & Russell, S. (2024). When code isn't law: Rethinking regulation for artificial intelligence. Policy and Society. https://doi.org/10.1093/polsoc/puae020

[16] Leach, T., et al. (2024). Trust, trustworthiness and AI governance. Scientific Reports, 14, Article 71761. https://doi.org/10.1038/s41598-024-71761-0

[17] Makridis, C. A. (2023). Artificial intelligence governance in companies: A systematic literature review and research agenda. Data & Policy, 5, e14. https://doi.org/10.1017/dap.2023.14

[18] Meijerink, J., et al. (2025). From principles to practice: Operationalizing AI governance in organizations. Business & Information Systems Engineering. https://doi.org/10.1007/s12599-024-00990-x

[19] Papagiannidis, S., Mikalef, P., & Conboy, K. (2025). Responsible AI governance: A review and research agenda. Journal of Strategic Information Systems. https://doi.org/10.1016/j.jsis.2024.101885

[20] Raji, I. D., Smart, A., White, J., & Mitchell, M. (2022). Operationalising AI governance through ethics-based auditing. AI & Society, 37(4), 1511–1529. https://doi.org/10.1007/s00146-021-01286-x

[21] Sadek, M., Kallina, E., Bohné, T., &Viglia, G. (2024). Challenges of responsible AI in practice: Scoping review and recommended actions. AI & Society, 40(1), 199–215. https://doi.org/10.1007/s00146-024-01880-9

[22] Salehi, M., Khaki, S., Amirkhani, A., & Naji, H. (2024). AI governance: A systematic literature review. AI and Ethics, 4, 1–37. https://doi.org/10.1007/s43681-024-00653-w

[23] Spiekermann, S., & Winkler, T. (2023). A collection of best practices for AI governance and engineering: The RAI pattern catalogue. ACM Computing Surveys. https://doi.org/10.1145/3626234

[24] Taeihagh, A. (2025). Governance of generative artificial intelligence. Policy and Society. https://doi.org/10.1093/polsoc/puaf001

[25] Zhou, Y., Parker, C., & He, K. (2022). Defining organizational AI governance. AI and Ethics, 2, 613–627. https://doi.org/10.1007/s43681-022-00143-x