

A literature Review on AI Governance and Human Rights

Olawunmi Azeezat Akanbi¹, Merveils Kadjo Shop², Hamna Seyyed³
^{1,2,3}Bahçeşehir Cyprus University

0009-0007-0887-0640 [Akanbi Olawunmi Azeezat]

0009-0008-1437-2776 [MERVEILS Kadjoshop]

Abstract

The exponential diffusion of Artificial Intelligence (AI) into the critical infrastructure of contemporary society including healthcare, judicial governance, and education has precipitated a profound normative crisis. As autonomous systems increasingly mediate fundamental human experiences, a systemic tension has emerged between algorithmic efficiency and the preservation of human rights. This review critically examines the necessary transition from "ethics-washing" characterized by voluntary, non-binding guidelines toward enforceable legal standards and robust accountability architectures. The primary objective of this study is to analyze emerging governance frameworks, specifically the Three-Layered Oversight model and Computational Constitutional Architecture (CCA), as mechanisms to bridge the "granularity gap" where traditional legal principles fail to address technical specificities. By synthesizing recent literature on "Control-by-Design" and Human Rights Impact Assessments (HRIA), the article evaluates how technocratic governance can be realigned with democratic imperatives. The findings suggest that the shift from corporate self-regulation to specialized national supervisory bodies and hard-coded legal protections is non-negotiable for the protection of human dignity, autonomy, and equality in the algorithmic era.

Keywords: Algorithmic Bias, Transparency, Privacy Protection, Responsible AI, Digital Governance

1. Introduction: The AI Frontier and Human Rights Imperatives

The transition of Artificial Intelligence from a peripheral technological novelty to the foundational architecture of critical infrastructure represents one of the most significant shifts in modern jurisprudence. The proliferation of these systems across healthcare, public governance, and education is no longer a speculative horizon but a pervasive reality (Cath, 2018). In the medical sector, AI mediates diagnostic procedures and resource allocation; in education, it shapes pedagogical trajectories and student surveillance; and in governance, it increasingly informs judicial decision-making and predictive policing (Kaliraman & Dhand, 2025). However, this rapid integration is often driven by a technocratic paternalism that prioritizes operational optimization over the nuanced protection of fundamental human rights. The primary challenge to established legal orders is the "Black Box" phenomenon—a state of ontological opacity where the internal logic of sophisticated, autonomous systems becomes inaccessible to human developers and stakeholders (Oyekunle et al., 2024). This technical opacity disrupts traditional notions of due process and the "right to an explanation." When algorithmic systems operate with high degrees of autonomy, the causal chain of decision-making is obscured, making the attribution of legal and ethical accountability increasingly elusive. This shift from human discretion to algorithmic scoring threatens to reduce the human subject to a mere data point, fundamentally undermining the principle of human dignity. The deployment of these systems introduces multifaceted risks to the core tenets of democratic societies:

- **Erosion of Autonomy and Dignity:** Automated decision-making systems often bypass individual agency, nudging behavior through choice architectures that compromise the capacity for independent self-determination (Chakraborty, 2025).
- **Algorithmic Bias and Structural Inequality:** Rather than eliminating human error, AI systems frequently formalize and amplify historical biases, leading to discriminatory outcomes in lending, employment, and law enforcement that disproportionately affect marginalized populations (Kunjukrishnan & Krishnakumariamamma, 2025).
- **Surveillance and Privacy Infringement:** The requirement for mass data ingestion to train large-scale models necessitates a regime of pervasive surveillance that erodes the private sphere and chills the exercise of civil liberties (Donahoe & Metzger, 2019).
- **Democratic Fragility:** The lack of transparency in algorithmic curation and the potential for automated manipulation of public discourse threaten the integrity of democratic processes and public trust in institutions. The central problem identified in contemporary scholarship is the "granularity gap"—the disconnect between high-level international human rights norms and the granular, technical execution of algorithmic code (Omosoto & Inedia, 2024). While the Universal Declaration of Human Rights provides a foundational ethos, it lacks the specificity required to govern the technical minutiae of algorithmic bias or data protection mandates. Consequently, there is an urgent imperative to develop comprehensive regulatory mechanisms that translate abstract rights into enforceable technical requirements, ensuring that the advancement of AI does not occur in a normative vacuum (Chakravarty et al., 2025).

2. Literature Review: The Evolution of AI Accountability

State-Centric vs. Private Sector Models Traditional accountability frameworks in human rights law have historically identified the state as the primary duty-bearer. However, the current AI landscape—dominated by a small number of multinational corporations—has rendered this state-centric model insufficient. Scholars now advocate for the direct extension of human rights obligations to private sector actors who control the development and deployment of critical societal tools (Dennehy et al., 2022). Effective oversight must encompass the entire supply chain, ensuring that private entities are held accountable for the socio-political repercussions of their technologies, particularly when they perform functions that were traditionally the remit of the state (McGregor et al., 2019). **From Ethics to Law: The Normative Gap** There is a burgeoning critique of the reliance on non-binding ethical guidelines, a trend often dismissed as "ethics-washing." These voluntary commitments frequently act as a strategic diversion to stall mandatory regulation (Hickok, 2020). A significant "normative gap" persists because current norms are perceived as too vague to curb discriminatory algorithmic practices or provide pathways for legal redress (Kriebitz, 2024). The shift toward binding legal duties is essential to provide clear mechanisms for enforcement and to ensure that human rights protections are not subject to corporate whim (Khalaileh, 2023). **Vulnerability-Based Frames** To improve the precision of regulation, recent proposals suggest adopting "vulnerability-based frames." This methodology involves identifying specific risks inherent in system architectures, such as algorithmic opacity, cybersecurity flaws, and the reproduction of harmful gender stereotypes (Ayinla et al., 2024). By focusing on the vulnerabilities created by AI—rather than just the outcomes—policymakers can create targeted regulations that address the root technical causes of human rights violations.

3. Methodological Frameworks for Robust Governance

Policy-Driven Systems Intelligence Governance must serve as the bridge between high-level policy objectives and technical execution. Policy-driven systems intelligence involves the design of AI where operational logic is inherently aligned with legal mandates. This ensures

that a system's performance is monitored not merely for its statistical accuracy, but for its adherence to human rights standards. **Control-by-Design** The "Control-by-Design" methodology advocates for the integration of human rights safeguards directly into the development lifecycle. This approach rejects the notion of "add-on" ethics, requiring instead that designers embed transparency, fairness, and accountability into the logic and code of the system from inception (Hoxhaj et al., 2023). **Human Rights Impact Assessments (HRIA)**, a vital tool for substantive compliance is the HRIA, which ensures that evaluations are conducted at every stage of the system's existence, from data sourcing to real-world deployment.

Lifecycle Stage	Focus of Assessment	Objective
Development	Data sourcing, algorithmic logic, and training set diversity.	Mitigate power asymmetries and identify latent algorithmic-use risks before training (Solanki et al., 2022).
Deployment	Context-specific application and user-system interaction.	Identify specific algorithmic-use risks and impacts on protected groups and democratic discourse.
Monitoring	Real-world outcomes and iterative feedback loops.	Evaluate the efficacy of auditing frameworks and ensure continuous adherence to international standards (Mökander & Floridi, 2021).

4. Advanced Accountability Architectures

Three-Layered Oversight To resolve the granularity gap, a multi-tiered monitoring approach is required to ensure system outcomes remain within legal parameters (Wang, 2025):

1. **Operational Layer:** Real-time monitoring of system inputs and outputs to detect immediate technical failures or biased results at the code level.
 2. **Tactical Layer:** Conducted by internal compliance teams to ensure systems meet pre-defined ethical and legal benchmarks through periodic reviews.
 3. **Strategic Layer:** External oversight by independent bodies to evaluate long-term societal impacts and ensure alignment with evolving international human rights law.
- Computational Constitutional Architecture (CCA)** A primary challenge in global AI governance is "international regulatory arbitrage," where corporations relocate operations to jurisdictions with weak human rights protections to bypass stringent laws (Su, 2022; Wong, 2020). Computational Constitutional Architecture addresses this by "hard-coding" fundamental rights protections into the AI's logic layer. By making these protections a technical requirement for the system's functionality, the jurisdiction of the hardware or the developer becomes irrelevant; the system is technically incapable of violating the hard-coded constitutional norms, thereby neutralizing the incentive for regulatory arbitrage (Su, 2022).

5. Participatory Governance: 'Society-in-the-Loop'

Multi-Stakeholder Collaboration; Effective governance requires a "Society-in-the-Loop" model, involving researchers, policymakers, and civil society. This ensures that legal mandates for algorithmic transparency and data protection are informed by diverse perspectives, particularly those of marginalized groups (Nuredin, 2023). **Independent Supervisory Bodies;** National authorities must evolve into specialized supervisory bodies acting as impartial mediators (Teo, 2023). These bodies serve as repositories for complaints and provide a platform for redress for individuals affected by AI-driven decisions. They are essential for enforcing

compliance through standardized certification and audit processes, ensuring that accountability is not merely performative (Sigfrids et al., 2022). Bottom-Up Requirements and Case Studies; Abstract norms are best translated into granular requirements through sector-specific applications. In legal practice, for instance, AI must be governed by strict confidentiality and safety protocols to protect the integrity of the judicial process (Abuzir, 2025). Similarly, the requirements for AI in healthcare (prioritizing safety) differ from those in finance (prioritizing non-discrimination). These bottom-up requirements ensure that regulations remain practical and actionable for developers.

6. Discussion: Mitigating 'Accountability Theatre' and Global Norms

Critique of Self-Regulation; The current reliance on corporate self-regulation is increasingly categorized as "Accountability Theatre." Industry actors frequently use corporate lobbying to resist mandatory audits, citing "innovation risks" to circumvent external oversight (Ibitoye et al., 2025). However, voluntary commitments have proven insufficient to prevent systemic human rights violations (Schultz et al., 2024). Mandatory impact assessments and independent audits are required to transition toward a model of genuine accountability. External Verification Independent auditing frameworks, such as the "Trustworthy AI Assessment," are essential for ensuring public accountability (Birkstedt et al., 2023). These frameworks provide a structured methodology to evaluate impacts on protected groups and children, ensuring that risk analyses remain aligned with human rights rather than corporate profit (Ayling & Chapman, 2021). **Global Collaboration:** To maintain a level playing field, international standardized reporting and confidentiality protocols are necessary (Cheong, 2024). These protocols must balance the protection of intellectual property with the transparency required for public trust. Global collaboration, supported by established international instruments, is essential to ensure that AI development remains accountable to the global human community (Stahl et al., 2023).

7. Conclusion

The rapid diffusion of AI across high-risk sectors necessitates a fundamental paradigm shift. The transition from voluntary, non-binding ethical guidelines to formal accountability procedures is now a legal and moral imperative. By implementing specialized supervisory bodies, Three-Layered Oversight, and Computational Constitutional Architectures, societies can bridge the granularity gap. Protecting human rights in the AI era requires a substantive commitment to transparency and the establishment of independent enforcement mechanisms that ensure technology serves human dignity, autonomy, and equality.

References

1. **Ahmed, I. (2025).** Navigating ethics and risk in artificial intelligence applications within information technology: A systematic review. *American Journal of Advanced Technology and Engineering Solutions*, 1(1), 579–601.
2. **Agarwal, A. K., & Nene, M. J. (2025).** A five-layer framework for AI governance: Integrating regulation, standards, and certification. *Transforming Government: People, Process and Policy*, 19(3), 535–555.
3. **Alberto, N. (2024).** XAI and Sustainability: Unifying Regulatory Standards and Solutions for the Future of Environmental Management. *EarthArXiv (California Digital Library)*.
4. **Akinrinola, O., Okoye, C. C., Ofodile, O. C., & Ugochukwu, C. E. (2024).** Navigating and reviewing ethical dilemmas in AI development: Strategies for transparency, fairness, and accountability. *GSC Advanced Research and Reviews*, 18(3), 50–58.

5. **Alahmadi, D. (2025).** AI Ethics Principles for AI Developers in Academia: A Critical Assessment and Analysis. *TEM Journal*, 2514–2529.
6. **Abhulimen, A. O., & Ejike, O. G. (2024).** Ethical considerations in AI use for SMEs and supply chains: Current challenges and future directions. *International Journal of Applied Research in Social Sciences*, 6(8), 1653–1679.
7. **Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023).** AI governance: Themes, knowledge gaps and future agendas. *Internet Research*, 33(7), 133–167
8. **Blanchard, A., Thomas, C. E., & Taddeo, M. (2024).** Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance. *AI & Society*.
9. **Cheong, B. C. (2024).** Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics*, 6.
10. **Costanza-Chock, S., Raji, I. D., & Buolamwini, J. (2022).** Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 1571–1583.
11. **Du, J. (2025).** Toward Responsible and Beneficial AI: Comparing Regulatory and Guidance-Based Approaches—A Comprehensive Comparative Analysis of Artificial Intelligence Governance Frameworks across the European Union, United States, China, and IEEE. *arXiv (Cornell University)*.
12. **Falco, G., Shneiderman, B., et al. (2021).** Governing AI safety through independent audits. *Nature Machine Intelligence*, 3(7), 566–571.
13. **Hoxhaj, O., Halilaj, B., & Harizi, A. (2023).** Ethical implications and human rights violations in the age of artificial intelligence. *Balkan Social Science Review*, 22(22), 153–171.
14. **Konidena, B. K., Malaiyappan, J. N. A., & Tadimarri, A. (2024).** Ethical Considerations in the Development and Deployment of AI Systems. *European Journal of Technology*, 8(2), 41–53.
15. **Koripalli, M. (2025).** Trust, Transparency, and Ethics: A Framework for Sustainable LLM Integration in Enterprise Information Systems. *Journal of Information Systems Engineering & Management*, 10, 376–383.
16. **Memeti, E., et al. (2024).** Navigating Ethical Terrain: Artificial Intelligence and Human Rights in the Digital Age. *Journal of Natural Sciences and Mathematics of UT*, 9, 314–325.
17. **Mittelstadt, B. (2019).** Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501–507.
18. **Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023).** Auditing large language models: A three-layered approach. *AI and Ethics*, 4(4), 1085–1115.
19. **Nuredin, A. (2023).** The Legal Status of Artificial Intelligence and the Violation of Human Rights. *Deleted Journal*, 2(1), 7–28.
20. **Nwachukwu, P. S., Chima, O. K., & Okolo, C. H. (2025).** The artificial intelligence governance framework for finance: A control-by-design approach to algorithmic decision-making in accounting. *Finance & Accounting Research Journal*, 7(8), 350–379.
21. **Osasona, F., et al. (2024).** Reviewing the Ethical Implications of AI in Decision Making Processes. *International Journal of Management & Entrepreneurship Research*, 6(2), 322–335.
22. **Shneiderman, B. (2020).** Bridging the Gap Between Ethics and Practice. *ACM Transactions on Interactive Intelligent Systems*, 10(4), 1–31.
23. **Thomas, C. E., et al. (2024).** The case for a broader approach to AI assurance: Addressing “hidden” harms in the development of artificial intelligence. *AI & Society*.

24. **Zafar, A. (2024).** Balancing the scale: Navigating ethical and practical challenges of artificial intelligence (AI) integration in legal practices. *Discover Artificial Intelligence*, 4(1).