"EXPLAINABLE AI AND ALGORITHMIC DIPLOMACY: THE EL-MDS FRAMEWORK FOR TRANSPARENT AND EQUITABLE MULTILATERAL DECISION-MAKING"

Research Paper

Ahmed Jajere, SSBM, Geneva, Switzerland, ahmed.jajere@student.ssbm.ch

“Abstract”

The escalating integration of Artificial Intelligence into high-stakes global governance contexts, from refugee allocation to international trade agreements, engenders profound opacity, eroding stakeholder trust and perpetuating epistemic inequalities and data colonialism, particularly impacting the Global South. This study introduces the Explainability Layers and Legibility of Multilateral Decision Systems (EL-MDS) framework, a novel socio-technical architecture designed to enhance transparency, accountability, and legitimacy in AI-driven multilateral decision-making. Employing a convergent parallel mixed-methods design, which integrates quantitative simulation experiments with qualitative analysis of secondary data and expert insights, the research evaluates the impact of EL-MDS on stakeholder trust and policy utility. Results reveal EL-MDS significantly improves perceived legitimacy and practical applicability of AI outputs, demonstrating its capacity to sustain higher trust levels even at elevated transparency. Crucially, ethical AI practices enhance trust only when mediated by robust organizational capabilities. EL-MDS offers a transferable governance blueprint for the ethical integration of AI and operationalizes "algorithmic diplomacy," fostering equitable global AI policy and combating the Digital Cantillon Effect.

Keywords: Explainable Artificial Intelligence, Algorithmic Diplomacy, Multilateral Governance, Transparency, Equity, AI Governance Framework, Institutional Trust.

1 Research Scope

This study aims to investigate and operationalize mechanisms that make AI-driven multilateral decisions legible, contestable, and equitable. The specific objectives are:

(i) To evaluate empirically (via ethically synthesized simulation data and qualitative policy analysis) the impact of EL-MDS on Stakeholder Trust and Acceptance (SCA), policy relevance, and policy applicability.

(ii) To assess the role of organizational capabilities and external institutional pressures as mediators of the relationship between ethical AI practices and stakeholder trust.

(iii) To elaborate policy and operational recommendations for piloting EL-MDS in multilateral settings, thereby countering the Digital Cantillon Effect and enabling algorithmic diplomacy.

1.1 Theoretical framework

This study’s theoretical grounding synthesizes literature on epistemic justice, explainable artificial intelligence (XAI), legibility, organizational trust, and the political economy of transnational governance. Rather than treating these literatures as parallel conversations, the framework stages them as mutually constitutive: technical explanations (XAI) do not produce legitimacy in isolation but must be rendered legible within institutional practices and power relations; organizational capacities and cognitive constraints shape whether explanations become usable; and multilateral norms and bargaining

dynamics determine who benefits from algorithmic opacity or transparency. Together, these strands motivate EL-MDS as a socio-technical governance architecture that is at once diagnostic (how models reason), mediative (how plural actors justify decisions), and procedural (how decisions can be contested and remediated). The following sub-sections unpack each strand and show how they generate testable hypotheses about the mechanisms through which EL-MDS can increase procedural justice and stakeholder trust.

1.1.1 Explainability, legibility, and epistemic justice

Explainability in machine learning research often focuses narrowly on producing technical artefacts — feature attributions, counterfactuals, or surrogate models —that reveal aspects of a model's internal logic. Yet, as philosophical work on epistemic injustice shows, the mere availability of information does not guarantee epistemic parity. Opacity in algorithmic systems can systematically deprive particular actors of the capacity to understand, interpret, or contest decisions, thereby entrenching forms of testimonial and hermeneutic injustice (Fricker, 2007). Building on this insight, recent work in critical XAI argues that technical explanations must be judged not only by fidelity or completeness but by their usability and cultural intelligibility (Kay, Kasirzadeh, and Mohamed, 2024).

EL-MDS reframes explainability as socio-technical legibility: an explanatory output is only meaningful when it is traceable (linked to auditable evidence), translatable (adapted to the epistemic practices and languages of diverse stakeholders), and contestable (embedded in institutional pathways for recourse). This triad traceability, translation, contestability foregrounds procedural norms: explanations should enable stakeholders to reconstruct reasons, surface normative premises, and, if necessary, trigger corrective processes. Importantly, this perspective shifts the evaluation criteria for XAI away from purely algorithmic metrics toward mixed evaluative rubrics that include cognitive accessibility, representational adequacy across cultures, and the capacity to support informed dissent (Mollema, 2024). In practice, this requires coupling technical diagnostics (e.g., SHAP/LIME outputs) with interpretive layers, simplified metaphors, domain-specific visualizations, and multilingual narratives and with institutional commitments to preserve explanation artifacts for audit and appeal.

1.1.2 Algorithmic diplomacy and the digital cantillon effect

Uneven distributions of design expertise, data access, and interpretive privilege characterize the political economy of algorithmic deployment. Drawing an analogy to Cantillon effects in monetary economies, we label the phenomenon whereby proximity to model design confers disproportionate advantage the Digital Cantillon Effect (Jajere, this study; Willson, 2025). Those actors who shape training datasets, tune model objectives, or control interpretive infrastructures acquire asymmetrical capacity to frame problems, select performance trade-offs, and capture downstream benefits. Left unaddressed, these asymmetries perpetuate and exacerbate geopolitical and epistemic inequalities within intergovernmental and regional institutions.

Algorithmic diplomacy reframes governance as an exercise in negotiated technical-political design: instead of unilateral standard-setting, states and institutions engage in co-design of interpretability thresholds, justification protocols, and shared audit standards that distribute epistemic authority more equitably (African Union, 2024; ASEAN, 2024). EL-MDS operationalizes algorithmic diplomacy by institutionalizing mechanisms such as a Federated Justification Framework (FJF) that aggregates contextualized rationales from a plurality of actor's ministries, civil society, and regional bodies, and encodes them alongside model outputs. The design logic is deliberate: by making plural justifications structurally visible and weighable, FJF reduces the capacity of single epistemic centers to monopolize decision narratives, thereby attenuating Digital Cantillon dynamics. The conceptual move here is to view explainability not as purely technical affordance but as a bargaining good that can be institutionalized to rebalance informational rents.

1.1.3 Organizational capabilities, cognitive load, and trust dynamics

Transparency is commonly presumed to be normatively unambiguously good; yet empirical and cognitive accounts indicate a more contingent relationship between transparency and trust. High-fidelity

disclosures may increase comprehension for some audiences while imposing prohibitive cognitive loads for others, resulting in confusion or misplaced skepticism (WTO, 2023). Parallelly, organizational readiness — including digital literacy, governance maturity, and procedural capacity determines whether institutions can operationalize ethical AI practices into reliable outcomes (IBM, 2023). Consequently, the effect of explainability on stakeholder trust is expected to be mediated and moderated by both cognitive and organizational factors.

EL-MDS therefore embeds testable assumptions about these causal pathways: (a) cognitive accessibility moderates the transparency trust relation such that beyond a certain threshold of raw information, additional detail yields diminishing or negative returns unless accompanied by interpretive scaffolding; (b) organizational capabilities mediate the impact of ethical AI practices on perceived legitimacy because only institutions with sufficient absorptive capacity can translate technical transparency into accountable policy actions; and (c) external institutional pressures (regulatory scrutiny, normative conditioning from peer institutions) amplify the credibility gains from explainability by creating incentives for follow-through (OECD, 2025). Methodologically, these claims invite mixed-methods operationalization: measurable constructs such as digital literacy indices, governance maturity scales, and experimentally manipulated explanation complexity will permit estimation of mediation and moderation effects in simulation and field settings.

1.1.4 Gaps and further research

Although XAI research has matured rapidly, the literature remains fragmented along disciplinary lines: computer science develops explanation algorithms and fidelity metrics; social sciences document interpretive challenges and normative risks; policy work proposes high-level standards. What is missing is an integrated governance architecture that can be instantiated across diverse multilateral settings, meaningfully translated across cultures, and empirically evaluated for both procedural justice and policy utility. In particular, three gaps motivate the EL-MDS research agenda: (1) scalability how to operationalize explanation-and-appeal workflows across institutions with varying capacities; (2) cross-cultural legibility how to measure and design explanations that maintain fidelity while being interpretable in multilingual and plural epistemic contexts; and (3) contestability how to create lightweight but effective recourse mechanisms that do not themselves become capture points.

Addressing these gaps requires a program of comparative, mixed-method research: multi-site pilots that pair technical modules with co-design workshops; ethnographic studies of decision workflows within target institutions; experimental manipulations of explanation form and complexity; and longitudinal evaluation of trust trajectories and policy outcomes. Such a program would also need to develop new metrics, e.g., a Legibility-Usability Index, a Federated Justification Diversity score, and organizational absorptive capacity scales that make governance trade-offs empirically tractable. EL-MDS contributes to this agenda by proposing a concrete modular architecture (DTE, FJF, CIL, AAI) whose components can be separately validated and iteratively refined through pilot deployments and stakeholder-driven evaluation, thereby translating normative ambitions for epistemic justice into operational design choices.

1.2 Research objectives

The study addresses the following primary research questions:

(i) To what extent does EL-MDS improve Stakeholder Trust and Acceptance (SCA) compared with baseline XAI?

(ii) (ii) How do organizational capabilities (e.g., digital literacy, governance structures) and external institutional pressures mediate or moderate the relationship between ethical AI practices and trust?

(iii) (iii) Can federated, justified, and appeal mechanisms reduce the Digital Cantillon Effect and produce more procedurally just outcomes?

(iv) (iv) What design principles and piloting strategies are necessary for scalable EL-MDS adoption in multilateral settings?

1.3 Research approach

This study adopts a convergent parallel mixed-methods design, a strategy particularly well suited to research questions that straddle both the technical evaluation of socio-technical systems and the normative assessment of institutional governance. The convergent design allows quantitative and qualitative strands to be developed simultaneously, with findings subsequently integrated to provide a richer and more triangulated understanding of how the Explainability Layers for Multilateral Decision Systems (EL-MDS) framework performs in contexts of institutional complexity and deep uncertainty. This methodological choice reflects the recognition that AI governance is not merely a computational or organizational issue but a hybrid domain where technical affordances, institutional capabilities, and cultural norms must be studied in concert.

On the quantitative side, the research employed ethically synthesized simulation data (n = 395) to recreate decision-making environments across multilateral contexts such as the OECD, ASEAN, ECOWAS, UNHCR, IMF, and NATO. The use of synthesized data served a dual purpose. First, it provided the statistical power and experimental control necessary to model complex relationships between transparency, organizational capacity, institutional pressures, and stakeholder trust. Second, it avoided the ethical and political sensitivities associated with analyzing confidential or politically charged decision data, while still ensuring the ecological plausibility of the simulated scenarios. Within this strand, multiple analytical techniques were deployed: descriptive statistics for baseline profiling, multivariate regression for testing explanatory power, mediation and moderation analyses for unpacking indirect pathways, and quadratic modeling for exploring hypothesized non-linear dynamics between transparency and trust. Through structuring the simulations around measurable constructs such as Stakeholder Trust and Acceptance (SCA), policy relevance, and policy applicability, the quantitative strand created a robust platform for stress-testing EL-MDS under varied institutional and governance conditions.

In parallel, the qualitative strand involved a systematic thematic analysis of 25 institutional and policy documents drawn from global, regional, and sectoral organizations engaged in AI governance. This included white papers, ethical frameworks, digital strategies, and regulatory guidelines produced by institutions such as the OECD, WTO, World Bank, African Union, and European Union. The selection of documents was guided by theoretical sampling: sources were chosen to reflect variation in institutional maturity, geographic representation, and governance models, ensuring that the analysis captured both convergent and divergent perspectives. The documents were coded deductively, based on the four EL-MDS modules (Decision Transparency Engine, Federated Justification Framework, Cross-Cultural Interpretability Layer, and Appeal and Audit Interface), and inductively, to capture emergent themes such as normative alignment, cognitive accessibility, and institutional legitimacy. This approach grounded the framework in real-world discourses and ensured that the simulated variables reflected empirically relevant governance concerns.

A third, integrative element of the research design was expert synthesis, where preliminary results were compared against insights from domain experts in AI ethics, organizational governance, and international relations. While informal in scope, this expert input played a crucial role in refining the operationalization of variables and validating the ecological plausibility of simulation parameters. Such synthesis helped bridge the gap between theoretical constructions and the practical realities of institutional decision-making, ensuring that the study's findings would resonate with both academic and policy audiences.

The integration of these three components, quantitative simulations, qualitative policy analysis, and expert synthesis, embodies the logic of methodological triangulation. Each strand addresses the limitations of the others: simulations provide causal leverage but risk abstraction; policy analysis grounds construct in institutional realities but cannot establish causal inference; and expert synthesis contextualizes findings while mitigating the risk of theoretical insularity. When converged, these strands

provide a holistic assessment of EL-MDS's capacity to enhance transparency, trust, and contestability in multilateral governance. This approach thus balances empirical rigor with normative sensitivity, making it a strong foundation for evaluating a socio-technical governance architecture intended for global institutional application.

1.3.1 EL-MDS architecture (core modules)

The EL-MDS framework comprises four interconnected modules:

(i) Decision Transparency Engine (DTE) produces model diagnostics, local/global explanations, and traceability reports (e.g., LIME/SHAP diagnostics and explanation dashboards), making model reasoning explicit and auditable.
(ii) Federated Justification Framework (FJF) aggregates weighted, context-specific justifications from institutional actors (ministries, NGOs, regional bodies), reducing dominance by single epistemic centers and embedding pluralist rationales in decision outputs.
(iii) Cross-Cultural Interpretability Layer (CIL) translates explanations across languages, political cultures, and epistemic traditions via simplified metaphors, regional visualizations, and multilingual outputs.
(iv) Appeal and Audit Interface (AAI) enacts structured recourse: time-stamped audit logs, appeal workflows, and recalculation mechanisms to enable contestability and accountability.

1.3.2 Quantitative simulations and measures

To address the methodological challenge of limited direct access to multilateral decision-making data, the study employed ethically synthesized datasets that replicated key institutional environments such as the OECD, ASEAN, ECOWAS, UNHCR, IMF, and NATO. These datasets were constructed to capture realistic decision variables, stakeholder compositions, and contextual asymmetries while avoiding ethical and political risks associated with the use of sensitive or proprietary data. The simulation environment provided a controlled yet sufficiently complex setting in which the performance of the EL-MDS framework could be stress-tested across varying governance and organizational capacity scenarios.

The central dependent variable was Stakeholder Trust and Acceptance (SCA), operationalized through multidimensional survey-based constructs capturing perceived transparency, fairness, legitimacy, and willingness to adopt AI-supported decision outputs. Additional dependent variables included policy relevance (the extent to which AI outputs aligned with institutional objectives) and policy applicability (the perceived feasibility of implementing recommendations in real-world governance contexts). Independent variables included measures of institutional transparency, ethical AI practices, organizational capability scales (capturing digital literacy, governance maturity, and absorptive capacity), and external institutional pressures.

Analytical techniques combined descriptive statistics (to profile stakeholder distributions and baseline perceptions) with multivariate regression models (to test the influence of independent variables on SCA and policy outcomes). To capture indirect pathways, mediation analyses were used to examine whether organizational capabilities mediated the relationship between ethical AI practices and trust, while moderation models tested whether external pressures amplified or attenuated these effects. Finally, quadratic modeling was applied to explore the hypothesized inverted-U relationship between transparency and trust, thereby assessing the cognitive load hypothesis that excessive disclosure may reduce perceived legitimacy if explanations become overwhelming or unintelligible.

The quantitative simulations thus provided a rigorous and replicable means of testing the causal architecture of EL-MDS, offering empirical evidence that complemented the qualitative findings. Importantly, this approach allowed the study to balance generalizability with contextual realism: the synthesized nature of the data enabled experimental manipulation of variables (e.g., transparency levels, justification mechanisms) that would be ethically or practically infeasible in live institutional settings, while still reflecting the dynamics of real-world multilateral governance.

1.3.3 Qualitative policy analysis

In parallel with the simulations, the study undertook a thematic analysis of 25 policy and institutional reports drawn from a diverse set of organizations engaged in AI governance, digital ethics, and multilateral decision-making. Sources included strategy documents, governance frameworks, and institutional white papers from the OECD, WTO, UNHCR, World Bank, African Union, ASEAN, and European Union, among others. The aim of this qualitative strand was not only to contextualize the simulation results but also to ensure that the design of EL-MDS aligned with the discursive priorities, normative expectations, and operational constraints identified in real-world policy environments.

The analysis followed a multi-stage coding process. First, a deductive coding framework was developed based on the core EL-MDS modules (Decision Transparency Engine, Federated Justification Framework, Cross-Cultural Interpretability Layer, and Appeal and Audit Interface). Reports were then coded for explicit references to transparency, accountability, justification, interpretability, and contestability. Second, inductive thematic coding identified emergent themes not fully captured in the initial framework, including repeated emphasis on norm alignment, cognitive accessibility, capacity asymmetries, and institutional legitimacy. The iterative process ensured that both theoretically anticipated and contextually emergent concerns were captured.

Findings revealed a strong convergence around five thematic priorities: (i) the demand for transparency that is actionable rather than symbolic; (ii) the necessity of norm alignment, ensuring AI systems respect institutional values and human rights commitments; (iii) the centrality of accountability mechanisms that make decisions auditable and contestable; (iv) concerns with cognitive load, highlighting the risk that overly complex explanations undermine rather than build trust; and (v) the importance of contextualization, emphasizing that AI governance must reflect local institutional cultures and capacities rather than relying on universalist templates.

These insights served two key purposes. First, they validated the constructs embedded in the simulation design by demonstrating their salience in real-world governance discourses. Second, they informed the questionnaire design for subsequent pilots, ensuring that survey items captured not only abstract measures of trust but also institutionally relevant concerns such as contestability, cognitive accessibility, and contextual fit. In doing so, the qualitative analysis grounded the EL-MDS framework in the lived realities of policy practice, thereby enhancing both the ecological validity of the study and the practical applicability of its recommendations.

1.4 Empirical findings (Summary of simulation results and thematic reinforcement)

The empirical findings present strong evidence that EL-MDS outperforms baseline XAI systems across trust, transparency, and policy applicability measures. Both quantitative simulations and thematic analyses reinforce the framework's ability to generate context-sensitive, actionable, and trusted outputs.

(i) Trust and Acceptance (SCA):
   a. EL-MDS simulations produced higher mean SCA ($M = 3.2$, $SD = 0.7$) compared to Baseline XAI ($M = 2.9$, $SD = 0.8$).
   b. A multivariate regression explained 49.2% of the variance ($R^2 = 0.492$, $F = 52.7$, $p < 0.001$).
   c. Organizational capabilities (CES) and external institutional pressures (EIP) were strong positive predictors of SCA.
   d. Ethical AI practices (GIP) had a weak direct effect but a significant indirect effect via CES, showing a mediating relationship.

(ii) Transparency and Trust (Inverted-U):
   a. Quadratic modeling revealed an inverted-U relationship between transparency and trust.

b. Trust peaked near ~80/100 on the OECD transparency index, beyond which cognitive overload reduced trust (WTO, 2023).
c. This highlights the need to pair high transparency with legible, user-friendly explanations.

(iii) Policy Relevance and Applicability:

a. EL-MDS significantly outperformed Baseline XAI in policy relevance (M = 4.03 vs 3.05) and applicability (M = 4.10 vs 2.88).
b. Effect sizes were very large (Cohen's d ≈ 1.85 and 2.31), confirming EL-MDS's ability to translate AI outputs into actionable policy recommendations.

(iv) Qualitative Alignment:

a. Thematic analysis emphasized the need for:
    i. Context-specific explanations
    ii. Norm alignment
    iii. Contestability mechanisms
b. Each of these themes maps to EL-MDS components (DTE, FJF, AAI, CIL).

1.5 Expected findings and practical implications

The expected findings extend the empirical results by showing how EL-MDS can advance theory, improve institutional practice, and inform policy interventions. Together, these implications form the foundation for testing EL-MDS in real-world multilateral decision-making contexts.

(i) Theory:

a. EL-MDS refines XAI theory by embedding explainability within institutional and cultural contexts.
b. It bridges technical diagnostics with procedural justice and epistemic fairness (Fricker, 2007; Mollema, 2024).

(ii) Practice:

a. EL-MDS offers concrete design patterns for multilateral institutions, including:
    i. Federated justification layers
    ii. Cross-cultural translation protocols
    iii. Appeal workflows
b. These patterns improve policy utility and stakeholder trust when combined with organizational capacity-building programs (IBM, 2023; OECD, 2025).

(iii) Policy:

a. Recommended steps include:
    i. Piloting EL-MDS through MOUs with partner institutions
    ii. Hosting co-design workshops
    iii. Conducting small-scale trials on single decision cycles (e.g., refugee quota allocation)
    iv. Running longitudinal evaluations to measure trust trajectories
b. These steps help assess scalability and tailor modules for lower-capacity institutions.

1.6 Conclusion

EL-MDS offers a transferable governance blueprint that enhances the legibility, contestability, and equity of AI-supported multilateral decision-making. Thereby integrating transparency diagnostics, federated justifications, cross-cultural interpretability, and appeal mechanisms, the framework moves beyond technical explainability toward a socio-technical model of algorithmic diplomacy. The simulation findings, supported by thematic policy analysis, confirm that stakeholder trust improves when transparency is paired with institutional capacity-building and when explanations are usable across

cultural and political contexts. This underscores the need to approach explainability not as an isolated technical task, but as a structural condition for epistemic justice and legitimacy in global governance.

At the same time, the study highlights the critical role of organizational readiness and external institutional pressures in shaping the effectiveness of ethical AI practices. Trust in multilateral AI systems is not generated automatically through transparency but requires meaningful institutional embedding, cognitive accessibility, and structured recourse. EL-MDS therefore contributes both a conceptual advance, mitigating the Digital Cantillon Effect and embedding algorithmic diplomacy, and a practical roadmap for pilot deployment. Future work should prioritize real-world testing, co-design workshops with diverse stakeholders, and iterative refinement to ensure scalability across institutions with varying capacities.

References

Adadi, A. and Berrada, M. (2020) 'Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information Fusion*, 58, pp. 82–115.

Ananny, M. and Crawford, K. (2018) 'Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability', *New Media and Society*, 20(3), pp. 973–989.

ASEAN Secretariat (2020) *ASEAN trade protocols white paper*. Jakarta: ASEAN Secretariat. Available at: https://asean.org (Accessed: 16 September 2025).

Baron, R.M. and Kenny, D.A. (1986) 'The moderator–mediator variable distinction in social psychological research', *Journal of Personality and Social Psychology*, 51(6), pp. 1173–1182.

Bianchi, I. (2024) 'Human rights and algorithmic accountability: Building effective governance structures for fair AI systems', *AI and Society*, 39(4), pp. 109–128.

Binns, R. (2018) 'Fairness in machine learning: Lessons from political philosophy', *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*, pp. 149–159.

Binns, R. and Veale, M. (2021) 'Is that your final decision? Multi-stage profiling, selective effects, and Article 22 of the GDPR', *International Data Privacy Law*, 11(4), pp. 319–332.

Bovens, M. (2007) 'Analysing and assessing accountability: A conceptual framework', *European Law Journal*, 13(4), pp. 447–468.

Burrell, J. (2016) 'How the machine "thinks": Understanding opacity in machine learning algorithms', *Big Data and Society*, 3(1), pp. 1–12.

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

Crawford, K. and Paglen, T. (2019) *Excavating AI: The politics of images in machine learning training sets*. New York: AI Now Institute.

Darwin, C.R. (1859) *On the origin of species by means of natural selection*. Cambridge: Harvard University Press (facsimile edn).

Doshi-Velez, F. and Kim, B. (2017) 'Towards a rigorous science of interpretable machine learning', *arXiv preprint* arXiv:1702.08608.

European Commission (2021) *Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Brussels: European Commission. Available at: https://commission.europa.eu (Accessed: 16 September 2025).

Finnemore, M. and Sikkink, K. (1998) 'International norm dynamics and political change', *International Organization*, 52(4), pp. 887–917.

Fishbein, M. and Ajzen, I. (1975) *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading, MA: Addison-Wesley.

Floridi, L. (2019) *The logic of information: A theory of philosophy as conceptual design*. Oxford: Oxford University Press.

Floridi, L. and Cowls, J. (2019) A unified framework of five principles for AI in society', *Harvard Data Science Review*, 1(1).

Gillespie, T. (2018) *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. New Haven: Yale University Press.

Hoff, K.A. and Bashir, M. (2015) 'Trust in automation: Integrating empirical evidence on factors that influence trust', *Human Factors*, 57(3), pp. 407–434.

Iyer, S.S. et al. (2024) 'The ethical implications of artificial intelligence in decision-making', Unpublished manuscript.

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399.

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G. and Yu, H. (2017) 'Accountable algorithms', *University of Pennsylvania Law Review*, 165(3), pp. 633–705.

Leslie, D. (2019) *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. London: Alan Turing Institute.

Lundberg, S.M. and Lee, S.I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30, pp. 4765–4774.

Miller, T. (2019) 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial Intelligence*, 267, pp. 1–38.

Mittelstadt, B.D. (2019) 'Principles alone cannot guarantee ethical AI', *Nature Machine Intelligence*, 1(11), pp. 501–507.

Mittelstadt, B.D., Allo, P., Taddeo, M., Wachter, S. and Floridi, L. (2016) 'The ethics of algorithms: Mapping the debate', *Big Data and Society*, 3(2), pp. 1–21.

Morley, J., Floridi, L., Kinsey, L. and Elhalal, A. (2020) '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*, 26, pp. 2141–2168.

Mudambi, R. and Navarra, P. (2004) 'Is knowledge power? Knowledge flows, subsidiary power and rent-seeking within MNEs', *Journal of International Business Studies*, 35(5), pp. 385–406.

Nisbett, R.E. et al. (2001) 'Culture and systems of thought: Holistic versus analytic cognition', *Psychological Review*, 108(2), pp. 291–310.

OECD (2021) *State of implementation of the OECD AI principles: Insights from national AI policies*. Paris: Organisation for Economic Co-operation and Development. Available at: https://www.oecd.org (Accessed: 16 September 2025).

Oxford Academic (2024) 'Governance of generative AI', *Policy and Society*, 44(1), pp. 1–20. Available at: https://academic.oup.com/policyandsociety/article/44/1/1/7997395 (Accessed: 16 September 2025).

PubMed (2025) 'Use of multi-criteria decision analysis (MCDA) to support decision-making during health emergencies: a scoping review', *Journal/Repository*, 2025. Available at: https://pubmed.ncbi.nlm.nih.gov/40416669/ (Accessed: 16 September 2025).

Raji, I.D. and Buolamwini, J. (2019) 'Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 429–435.

Rahwan, I. (2018) 'Society-in-the-loop: Programming the algorithmic social contract', *Ethics and Information Technology*, 20(1), pp. 5–14.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) '"Why should I trust you?": Explaining the predictions of any classifier', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144.

Rico, P. (2024) 'AI and data governance: A legal framework for algorithmic accountability and human rights', *International Journal of Law and Technology*, 22(2), pp. 133–152.

Royal Society (2020) *Explainable AI: The basics*. London: The Royal Society. Available at: https://royalsociety.org (Accessed: 16 September 2025).

Selbst, A.D. and Barocas, S. (2018) 'The intuitive appeal of explainable machines', *Fordham Law Review*, 87(3), pp. 1085–1139.

Shankar IAS Parliament (2025) 'UPSC daily current affairs: Prelimbits 04-08-2025', *Shankar IAS Parliament*, 4 August. Available at: https://www.shankariasparliament.com/blogs/pdf/upsc-daily-current-affairs-prelimbits-04-08-2025/ (Accessed: 16 September 2025).

Sidley Austin LLP (2023) 'Legal aid, free advice and virtue signalling', *Sidley Austin Insights*, January. Available at: https://www.sidley.com/en/insights/publications/2023/01/legal-aid-free-advice-and-virtue-signalling/ (Accessed: 16 September 2025).

Smith, J. and Jones, A. (2022) 'Translating technical explanations for policy audiences: best practices', *Journal of Public Policy and Technology*, 4(2), pp. 45–62.

Suresh, H. and Guttag, J. (2021) 'A framework for understanding sources of harm throughout the machine learning life cycle', *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 1–12.

UNDP (2022) *AI governance and human rights: A guide for policymakers*. New York: United Nations Development Programme. Available at: https://www.undp.org (Accessed: 16 September 2025).

United Nations (2021) *Roadmap for digital cooperation: Implementation report*. New York: United Nations. Available at: https://www.un.org (Accessed: 16 September 2025).

Vashistha, A. (2022) 'Making AI explainable in the Global South: A systematic review', XAI4D Compass. Available at: https://www.adityavashistha.com/uploads/2/0/8/0/20800650/xai4d-compass-2022.pdf (Accessed: 16 September 2025).

Wagner, B. (2022) 'Algorithmic diplomacy and the politics of AI standards', *Global Studies Quarterly*, 2(3), pp. 1–11.

WJARR (2025) 'Algorithmic bias, data ethics, and governance: Ensuring fairness', *World Journal of AI, Regulation and Rights*, 2025. Available at: https://journalwjarr.com/sites/default/files/fulltext_pdf/WJARR-2025-0571.pdf (Accessed: 16 September 2025).

World Bank (2023) *AI for development: Policy pathways for low-income countries*. Washington, DC: World Bank. Available at: https://www.worldbank.org (Accessed: 16 September 2025).

WTO (2024) *Integrating AI and blockchain for equitable trade*. Geneva: World Trade Organization.

Zerilli, J., Knott, A., Maclaurin, J. and Gavaghan, C. (2019) *Algorithmic decision-making and the law*. Oxford: Oxford University Press.