RAISE: A Responsible AI Implementation Framework through System Ethics for Governing Automated Decision-Making in Business Operations

Prince Nweke Onyeka
Data Analytics and Technologies
University of Greater Manchester, United Kingdom

Chinonso Job
School of Built Environment, Engineering, and Computing
Leeds Beckett University, United Kingdom

Festus Chijioke Onwe
Information Technology Department
University of Port Harcourt, Rivers State, Nigeria
Corresponding Author

ABSTRACT
A companion critical review identifies four structural governance gaps in the deployment of AI-driven Automated Decision-Making (AI-ADM) systems across business operations: the absence of systematic algorithmic auditing mechanisms, inadequate reskilling infrastructure for automation-displaced workers, the lack of standardised environmental lifecycle accounting for AI systems, and the complexity of cross-jurisdictional regulatory compliance. This paper proposes RAISE (Responsible AI Implementation through System Ethics), a five-pillar governance framework designed to address these gaps in an integrated and practically deployable manner. The five pillars are: (R) ethics-by-design, embedding ethical requirements into AI-ADM development from inception; (A) AI Ethics Governance Committees, providing multi-disciplinary oversight throughout the AI lifecycle; (I) human-in-the-loop (HITL) systems, maintaining meaningful human oversight in high-stakes decision paths; (S) Green AI practice, minimising the environmental footprint of AI operations; and (E) explainability and accountability mechanisms, ensuring interpretable, auditable, and contestable ADM outputs. RAISE is presented as a design-science artefact following Hevner et al.'s design-science research methodology: each pillar is grounded in published evidence and professional standards, the framework is specified at sufficient operational detail to support organisational implementation, and an evaluation protocol is provided for empirical testing. Feasibility constraints—including cost and resource implications, technical complexity, and organisational change management requirements—are assessed candidly for each pillar. The framework addresses the documented gap between high-level AI ethics principles, which are extensively published but weakly operationalised, and concrete, organisation-level governance practice.

Categories and Subject Descriptors
K.4.1 [Computers and Society]: Public Policy Issues — ethics; I.2.0 [Artificial Intelligence]: General — governance; H.1.2 [Models and Principles]: User/Machine Systems — human factors

General Terms
Design, Management, Human Factors, Standardization, Legal Aspects

Keywords
Responsible AI, AI governance framework, ethics-by-design, human-in-the-loop, Green AI, explainable AI, AI Ethics Committee, automated decision-making, AI accountability

1. INTRODUCTION
The governance of AI-driven Automated Decision-Making (AI-ADM) systems in business operations is characterised by a documented implementation gap: high-level ethical principles are widely published and broadly agreed upon in both academic and policy literature, while concrete, organisation-level mechanisms for operationalising those principles remain underdeveloped [1]. Jobin et al.'s analysis of 84 AI ethics guidelines across national and international organisations identifies a global convergence around five principles—transparency, fairness, non-maleficence, responsibility, and privacy—but finds persistent divergence in the extent to which these principles are translated into actionable practice [1]. The consequence, as documented in the companion critical review of AI-ADM implications, is that ethical commitments stated at the level of corporate AI principles are frequently not reflected in the design decisions, operational processes, or accountability structures of deployed AI-ADM systems.

This paper proposes the RAISE framework—Responsible AI Implementation through System Ethics—as a five-pillar governance artefact designed to close this implementation gap. RAISE is grounded in the design-science research (DSR) paradigm, as formalised by Hevner et al., in which knowledge and understanding of a problem domain are achieved through the building and evaluation of designed artefacts rather than through the verification of descriptive theories alone [2]. Consistent with DSR methodology, RAISE is: (i) specified at sufficient operational detail to be implemented by organisations deploying AI-ADM systems; (ii) grounded in a review of the published evidence base for each component; and (iii) accompanied by an evaluation protocol that defines the conditions under which the framework's central claims should be tested.

RAISE is structured around five pillars: (R) ethics-by-design; (A) AI Ethics Governance Committees; (I) human-in-the-loop systems; (S) Green AI practice; and (E) explainability and accountability mechanisms. Together, these pillars address the four cross-cutting gaps identified in the companion review: the audit and accountability gap (Pillars A and E); the reskilling infrastructure gap (Pillar A and I); the environmental accounting gap (Pillar S); and the cross-jurisdictional compliance gap (Pillar R and A). Section 2 motivates the framework; Section 3 specifies each pillar; Section 4 presents the evaluation protocol; Section 5 assesses feasibility and limitations; and Section 7 concludes.

2. BACKGROUND AND MOTIVATION
The need for a structured, multi-pillar governance framework for AI-ADM emerges from three convergent findings in the literature.

First, the operational consequences of ungoverned AI-ADM are documented and material. Raghavan et al.'s evaluation of algorithmic hiring tools found that vendor-disclosed bias-

mitigation practices were inconsistent and often insufficient to prevent discriminatory outcomes, even when organisations believed their systems were adequately governed [3]. The accountability gap this creates is real: when an AI-ADM system produces a biased or erroneous outcome, the diffuse architecture of most deployed systems—spanning data providers, model developers, platform operators, and organisational users—makes it structurally difficult to assign responsibility and ensure remediation [4].

Second, existing regulatory frameworks, while necessary, are structurally insufficient as standalone governance mechanisms. GDPR Article 22's right to explanation [5] and the EU AI Act's conformity assessment requirements for high-risk AI systems [6] establish legal obligations but do not specify operational procedures for meeting them. Organisations need not merely compliance checklists but an integrated governance architecture that embeds ethical requirements into development, operations, and oversight from inception.

Third, current AI ethics frameworks and principles, while widely cited, are insufficiently operationalised. Floridi and Cowls' five-principle framework—beneficence, non-maleficence, autonomy, justice, and explicability—provides a sound normative foundation [4]. Jobin et al.'s global landscape analysis confirms that these principles are reproduced, in various forms, across the majority of published AI ethics guidance [1]. Yet the operationalisation of these principles—the specific mechanisms, processes, and accountability structures needed to translate them from statement to practice—is precisely what is missing from the current governance landscape [7].

RAISE addresses this operationalisation gap. Each pillar is motivated by evidence of the specific mechanism it provides, specified at operational detail, and assessed for feasibility.

3. THE RAISE FRAMEWORK

Figure 1 illustrates the RAISE framework's five-pillar architecture and the governance lifecycle through which it operates.

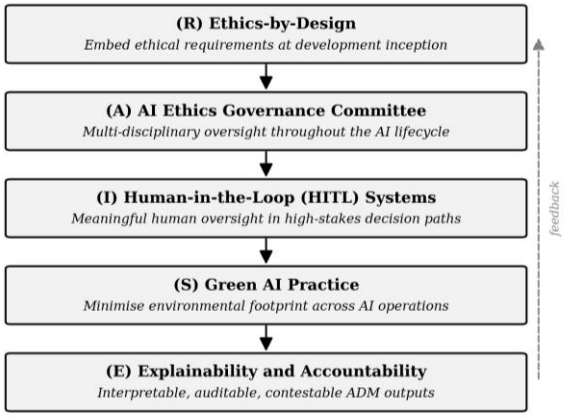

Fig. 1. The RAISE framework: five governance pillars operating as a sequential and iterative lifecycle, from system design (R) through ongoing oversight (A, I, S) to output accountability (E), with feedback from E informing the next design cycle.

A. Pillar R: Ethics-by-Design

Ethics-by-design operationalises the principle that ethical requirements must be embedded into AI-ADM systems at the earliest stage of the development lifecycle, before deployment decisions and technical architecture choices have foreclosed options [4]. This pillar is motivated by a well-documented failure mode: systems that are assessed for ethical adequacy only after deployment are constrained by existing architecture to post-hoc mitigations—adjustments to model outputs or decision thresholds—that are less effective than structural design choices made at inception [8].

Ethics-by-design encompasses three specific mechanisms.

The first is bias-aware data governance: training datasets for AI-ADM systems must be audited for demographic representativeness, temporal relevance, and the presence of proxy variables that could serve as discriminatory features. Raghavan et al.'s documentation of vendor bias-mitigation practices illustrates what an insufficient data governance process looks like: historical performance data used as a target variable for hiring models will encode prior discrimination if the workforce whose performance it reflects was itself selected through biased processes [3]. Bias-aware data governance requires active underrepresentation correction, proxy-variable identification and removal, and temporal data pruning to exclude periods of documented discriminatory practice.

The second mechanism is fairness-constrained model development. Barredo Arrieta et al.'s XAI taxonomy identifies model families with varying intrinsic interpretability levels [8]; ethics-by-design requires that the choice of model architecture reflect the interpretability requirements of the decision context. In high-stakes ADM contexts—where GDPR Article 22 obligations apply and the EU AI Act classifies the system as high-risk—model choices that prioritise predictive accuracy at the cost of interpretability are not appropriate without supplementary XAI mechanisms [5], [6].

The third mechanism is ethical impact assessment before deployment. Analogous to the Data Protection Impact Assessment required by GDPR Article 35 for high-risk data processing, an ethical impact assessment for AI-ADM should systematically evaluate: the potential for discriminatory outcomes across affected population subgroups; the reversibility of automated decisions and the availability of contestation mechanisms; and the scope of human oversight maintained in the operational decision flow [4], [9].

B. Pillar A: AI Ethics Governance Committee

An AI Ethics Governance Committee (AEGC) is a permanent, multi-disciplinary body responsible for continuous oversight of AI-ADM systems throughout their operational lifecycle. The committee's composition should span AI researchers and engineers, who can assess technical claims and risks; legal and compliance experts, who can evaluate regulatory obligations; domain specialists from the operational contexts in which AI-ADM is deployed (HR, finance, logistics); ethicists or social scientists with expertise in algorithmic fairness; and representatives of affected stakeholder groups [1], [7].

The AEGC's functions are operationally distinct from those of existing governance bodies such as data protection officers, audit committees, or information security teams, and should not be collapsed into them. Its specific responsibilities include:

• Reviewing and approving ethical impact assessments (Pillar R) before AI-ADM systems proceed to deployment;

• Establishing and maintaining an algorithmic auditing programme that regularly evaluates deployed systems for bias, accuracy, and fairness, using disaggregated performance metrics across demographic and operational subgroups;

• Setting and enforcing organisational standards for human

oversight in AI-ADM decision paths (linking to Pillar I);

• Receiving, investigating, and responding to individual contestations of AI-ADM decisions under GDPR Article 22 and DUAA 2025 rights [5], [10];

• Reporting findings and recommendations to senior leadership and, where relevant, to regulators.

The AEGC model is not novel—analogous bodies exist in financial risk management (model risk committees) and pharmaceutical development (independent ethics review boards)—but its application to AI-ADM governance in business operations has not been systematically operationalised. Kunkel et al.'s stakeholder involvement analysis supports the multi-stakeholder composition model, finding that inclusion of diverse perspectives—including affected community representatives—produces more robust identification of ethical risks than expert-only panels [7].

C. Pillar I: Human-in-the-Loop Systems

Human-in-the-loop (HITL) systems maintain meaningful human involvement in AI-ADM decision paths, ensuring that consequential decisions are either reviewed by, or reversible by, a human agent with the authority and contextual knowledge to exercise independent judgement. The specification of HITL is important: nominal human involvement—for instance, a human who clicks to execute an AI recommendation without any independent review—does not constitute meaningful oversight and should not be treated as satisfying HITL requirements [11].

RAISE specifies three operationally distinct HITL configurations, calibrated to decision stakes:

• **Human-out-of-the-Loop (HotL):** AI-ADM operates autonomously in the decision path, with humans monitoring outputs and retaining the authority to intervene when anomalies are detected. Appropriate for lower-stakes decisions (e.g. supply chain inventory replenishment) with well-understood error modes and reversible outcomes.

• **Human-in-the-Loop (HITL):** AI-ADM produces a recommendation or classification, which a human reviews before a consequential decision is executed. Appropriate for medium-stakes decisions (e.g. loan approval with standard parameters, recruitment shortlisting before interview scheduling).

• **Human-as-the-Loop (HatL):** AI-ADM provides analytical support—pattern identification, anomaly detection, risk scoring—but the decision itself is made by a human who is not bound by the AI output. Required for high-stakes, high-consequence decisions (e.g. decisions affecting individual employment, significant credit actions, safety-critical operations) where EU AI Act high-risk classification applies [6].

This tiered structure provides a practical operationalisation of the EU AI Act's human oversight requirements, which mandate that high-risk AI systems be designed and developed in a way that allows effective human oversight [6], without the impractical implication that all AI-ADM outputs require equivalent levels of human review regardless of stakes.

Google Cloud's human-in-the-loop documentation confirms that HITL configurations are practically deployed at scale across enterprise AI systems, providing operational precedent for the tiered model [12]. Gupta et al.'s empirical study of bias mitigation in large language models demonstrates that human feedback loops, when structured around specific fairness criteria rather than general approval, produce measurable bias reduction over successive model iterations [13].

D. Pillar S: Green AI Practice

Green AI practice addresses the environmental sustainability dimension of AI-ADM governance, seeking to minimise the energy consumption, greenhouse gas emissions, water use, and hardware waste generated by AI systems without forfeiting the operational benefits of AI-ADM. This pillar is motivated by Strubell et al.'s quantification of AI training costs [14] and Lannelongue et al.'s broader Green Algorithms framework [15], both of which establish that the environmental externality of large AI model deployment is material and growing.

RAISE identifies four specific Green AI mechanisms:

• **Model efficiency optimisation:** systematic evaluation of the accuracy/efficiency trade-off in model selection and design, prioritising architectures that achieve required decision quality with reduced computational overhead. Where simpler, lower-energy models achieve equivalent decision quality for the specific ADM context, they should be preferred over larger, higher-energy models that provide marginal accuracy gains [14].

• **Green data centre procurement:** preference in cloud service provider selection for facilities with verified renewable energy procurement and high Power Usage Effectiveness (PUE) ratings. Organisations should include environmental performance criteria in AI infrastructure procurement decisions, analogous to the environmental criteria now routinely included in supply chain procurement.

• **Lifecycle hardware management:** implementation of systematic AI hardware lifecycle management programmes, including extended hardware refresh cycles where technically feasible, certified e-waste disposal, and participation in manufacturer take-back programmes. The UN Global E-waste Monitor 2024's finding that only 22.3% of global e-waste is properly recycled [16] illustrates the extent of the current gap between disposal volumes and responsible disposal practice.

• **Environmental AI Impact Assessment:** analogous to the Ethical Impact Assessment in Pillar R, organisations should conduct a structured environmental assessment before deploying large AI-ADM systems, estimating expected energy consumption, emissions, and hardware lifecycle impacts and comparing them against organisational sustainability targets.

E. Pillar E: Explainability and Accountability Mechanisms

Pillar E operationalises transparency, explainability, and accountability requirements across deployed AI-ADM systems, addressing the XAI gap identified in the companion review [8]. Three mechanisms are specified:

• **Contextual XAI deployment:** selection and deployment of XAI methods appropriate to the decision context and the explanation audience. Barredo Arrieta et al.'s taxonomy distinguishes between global explanations (which characterise the overall behaviour of a model) and local explanations (which characterise a specific decision); between post-hoc explanations (which approximate complex model logic after the fact) and inherently interpretable models (which are transparent by design) [8]. RAISE requires that, for any AI-ADM system in scope, the deployed XAI method is specified, documented, and demonstrably adequate to support the right to explanation under GDPR Article 22 and DUAA 2025 [5], [10].

• **Individual contestation protocols:** a documented, time-bounded, and responsive process through which individuals who have been subject to an AI-ADM decision can request a human review, receive a meaningful explanation of the decision logic,

and have demonstrably incorrect decisions corrected and remediated. Kaminski's analysis of the right to explanation identifies the absence of such protocols as the primary practical gap in current GDPR compliance practice [9]; RAISE provides a template protocol that organisations can adapt to their specific ADM contexts.

• **Public AI system registers:** publication of summary documentation for deployed AI-ADM systems, specifying their purpose, the types of decisions they make, the data inputs they use, the performance metrics by which they are evaluated, and the oversight mechanisms maintained. This requirement is directly motivated by the EU AI Act's mandatory transparency obligations for high-risk AI systems [6] and extends to non-EU contexts as a governance best practice. Mitchell et al.'s model card framework provides a directly applicable technical template for such documentation [17].

4. EVALUATION PROTOCOL

Consistent with Hevner et al.'s design-science research principles [2], RAISE's claims should be tested against an explicit evaluation protocol. The framework proposes the following four-stage evaluation for any organisation implementing RAISE:

Stage 1 (Pre-implementation baseline): measure existing AI-ADM governance practice against four baseline indicators: (i) the proportion of deployed AI-ADM systems with documented ethical impact assessments; (ii) the disaggregated false-positive and false-negative rates of deployed ADM systems across demographic subgroups; (iii) the organisation's AI energy consumption and emission estimate; and (iv) the average time-to-resolution for individual contestation requests.

Stage 2 (Implementation): deploy each RAISE pillar in sequence over a 12-month implementation period, beginning with Pillar R (ethics-by-design for new AI-ADM developments) and Pillar A (AEGC establishment), followed by Pillars I, S, and E.

Stage 3 (12-month post-implementation assessment): remeasure the four baseline indicators and evaluate against three additional outcome metrics: user satisfaction ratings from employees and customers affected by AI-ADM decisions; the number and resolution rate of individual contestations under Pillar E protocols; and the energy and emissions reduction achieved under Pillar S.

Stage 4 (Longitudinal validation): conduct repeat assessment at 24 and 36 months to assess persistence of governance improvements and identify pillar-specific implementation challenges that the baseline evaluation period may not have captured. Longitudinal assessment is necessary because AI-ADM system populations within organisations change over time as new systems are deployed and existing ones are updated.

5. FEASIBILITY ASSESSMENT

A. Pillar Feasibility and Implementation Challenges

Each RAISE pillar carries specific feasibility considerations that organisations should assess against their context before implementation.

Pillar R (Ethics-by-Design) is most feasible for new AI-ADM systems, where design choices remain open; retrofitting ethics-by-design to legacy systems with established architectures is substantially more challenging and may require phased replacement rather than in-place modification. The bias-aware data governance requirement is resource-intensive where training data exists across multiple systems or external sources; organisations with limited data engineering capacity may require external specialist support.

Pillar A (AI Ethics Governance Committee) requires both structural resources (dedicated meeting capacity, administrative support) and multi-disciplinary expertise that many organisations will not hold internally. The inclusion of domain specialists from operational contexts (HR, finance, logistics) and affected stakeholder representatives may require the appointment of external advisors. A particular risk is committee capture—the tendency for multi-disciplinary governance bodies to be dominated by technical experts whose framing of issues reflects engineering rather than social or ethical priorities. Explicit terms of reference that mandate balanced representation and structured deliberation processes can mitigate this risk [7].

Pillar I (HITL Systems) involves a direct trade-off between decision speed and oversight quality. In ADM contexts with high decision volumes—automated fraud detection, real-time inventory management—HITL review of every decision is operationally infeasible. The tiered HotL/HITL/HatL specification addresses this by calibrating oversight intensity to decision stakes. However, determining the appropriate tier for each ADM context requires governance judgement that the AEGC (Pillar A) should provide. The risk of "automation bias"—the documented tendency for humans to defer to AI recommendations even when they have the authority and information to override them—requires specific mitigation through training and procedural design of HITL review processes [12].

Pillar S (Green AI) faces a significant data availability challenge: current AI energy consumption reporting is not standardised, making the baseline measurement required in Stage 1 of the evaluation protocol difficult to execute reliably. Lannelongue et al.'s Green Algorithms framework provides a practical starting point for estimation [15], but organisation-level reporting will require tooling investment. The Green AI requirement is also in potential tension with performance objectives: more energy-efficient models may be less accurate for specific ADM tasks, and organisations must manage this trade-off explicitly rather than assuming that sustainability and performance are always complementary.

Pillar E (Explainability and Accountability) is technically mature for certain model families—linear models, decision trees, and gradient boosted models with SHAP-based explanations are well supported by current XAI tooling—but remains challenging for deep neural network architectures, which are increasingly common in enterprise AI-ADM. Barredo Arrieta et al.'s taxonomy documents that the fidelity of post-hoc XAI approximations of deep model behaviour is limited and should be explicitly disclosed when used as the basis for explanation under GDPR [8].

B. Unsatisfactory Approaches and Counterpatterns

RAISE explicitly identifies three governance counterpatterns that organisations should avoid.

Ethics theatre: adoption of formal governance structures—AEGC formation, ethics-by-design policy publication, HITL labelling—that satisfy the form of governance requirements without the substance. Ethics theatre is common where governance investment is cost-constrained or where governance function is absorbed into existing compliance teams without dedicated resources.

Automation bias under HITL cover: maintaining nominal human-in-the-loop structures while organisational incentives and time pressures mean that human reviewers in practice approve

AI-ADM recommendations without meaningful independent evaluation. This pattern produces the regulatory appearance of human oversight while delivering the practical reality of fully automated decision-making [11].

Greenwashing of AI sustainability: publishing high-level green AI commitments—renewable energy pledges, net-zero targets—without operationalising the specific changes in model selection, infrastructure procurement, and hardware lifecycle management that would reduce actual environmental impact [15].

6. DISCUSSION: THE RAISE CONTRIBUTION

RAISE makes contributions at three levels. At the theoretical level, it advances the design-science research literature on AI governance by providing a structured, multi-pillar framework that is simultaneously grounded in empirical evidence, specified at operational detail, and accompanied by a falsifiable evaluation protocol—attributes that distinguish it from the high-level principles frameworks that dominate the existing AI ethics literature [1], [4].

At the regulatory level, RAISE provides a compliance architecture that operationalises the obligations imposed by GDPR Article 22 [5], the UK DUAA 2025 [10], and the EU AI Act [6] within a single integrated framework rather than treating each regulatory requirement as a separate compliance exercise. This integrated approach is more efficient than regulatory-silo compliance and reduces the risk of requirements being addressed in a piecemeal and potentially contradictory manner.

At the practical level, RAISE provides organisations—particularly those without dedicated AI ethics teams or established AI governance practice—with a structured roadmap for building responsible AI-ADM governance from the ground up. The pillar sequence (R A I S E) reflects a logical implementation priority: ethics-by-design and governance committee establishment should precede operational oversight and accountability mechanisms, because the latter depend on policy and oversight infrastructure that the former provides.

7. CONCLUSION AND FUTURE WORK

This paper has proposed RAISE, a five-pillar framework for governing AI-ADM systems in business operations, addressing four structural gaps in current governance practice. Ethics-by-design (R) addresses the training data and model selection deficiencies that generate bias and opacity at source. The AI Ethics Governance Committee (A) provides the multi-disciplinary, lifecycle-persistent oversight that organisational AI governance requires but rarely achieves. Human-in-the-loop systems (I) maintain meaningful oversight in high-stakes decision contexts without imposing operationally infeasible review burdens on high-volume, lower-stakes decision flows. Green AI practice (S) addresses the environmental externality of AI-ADM deployment through model efficiency, infrastructure procurement, and hardware lifecycle management. Explainability and accountability mechanisms (E) ensure that AI-ADM systems produce outputs that are interpretable, auditable, and contestable at the level required by applicable regulatory frameworks.

RAISE is presented as a conceptual design-science artefact: its empirical validation, through the evaluation protocol described in Section 4, remains the subject of future work. Three specific future research directions are identified. First, empirical validation across multiple business contexts (HR, financial services, supply chain) to determine whether the tiered HITL specification, and the allocation of AI-ADM contexts to specific tiers, is context-dependent in ways the current specification does not capture. Second, development of standardised environmental lifecycle accounting tooling that would make Pillar S's Environmental AI Impact Assessment practically executable across organisations of different technical maturity. Third, study of the AEGC's governance effectiveness across different committee composition models, to determine the optimal balance of technical, legal, and stakeholder expertise that produces the best governance outcomes across the diversity of AI-ADM deployment contexts.

8. STATEMENT ON VALIDATION STATUS

RAISE has not been implemented or empirically evaluated at the time of writing. All performance claims are prospective and conditional on the evaluation protocol described in Section 4 being executed. No claim in this paper should be read as an empirical result for the RAISE framework as deployed.

9. REFERENCES

[1] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nature Machine Intelligence, vol. 1, no. 9, pp. 389–399, 2019.

[2] A. R. Hevner, S. T. March, J. Park, and S. Ram, "Design science in information systems research," MIS Quarterly, vol. 28, no. 1, pp. 75–105, 2004.

[3] M. Raghavan, S. Barocas, J. Kleinberg, and K. Levy, "Mitigating bias in algorithmic hiring: Evaluating claims and practices," in Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), 2020, pp. 469–481.

[4] L. Floridi and J. Cowls, "A unified framework of five principles for AI in society," in Machine Learning and the City: Applications in Architecture and Urban Design. Cham: Springer, 2022, pp. 535–545.

[5] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 general data protection regulation (GDPR)," Official Journal of the European Union, L 119, 2016.

[6] "Regulation (EU) 2024/1689 artificial intelligence act," Official Journal of the European Union, 2024.

[7] S. Kunkel, F. Schmelzle, S. Niehoff, and G. Beier, "More sustainable artificial intelligence systems through stakeholder involvement?" GAIA – Ecological Perspectives for Science and Society, vol. 32, no. 1, pp. 64–70, 2023.

[8] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020.

[9] M. E. Kaminski, "The right to explanation, explained," in Research Handbook on Information Law and Governance. Cheltenham: Edward Elgar Publishing, 2021, pp. 278–299.

[10] UK Parliament, "Data (Use and Access) Act 2025," https://www.legislation.gov.uk/ukpga/2025/18, 2025.

[11] M. Khaleel, A. A. Ahmed, and A. Alsharif, "Artificial intelligence in engineering," Brilliance: Research of Artificial Intelligence, vol. 3, no. 1, pp. 32–42, 2023.

[12] Google Cloud, "What is human-in-the-loop (HITL) in AI

and ML," https://cloud.google.com/discover/human-in-the-loop, 2024, accessed 27 June 2025.

[13] S. Gupta, R. Kundu, A. K. A. Deo, M. Patnaik, T. Kundu, and M. K. Dehury, "Enhancing transparency and mitigating bias in large language models' responses with sophistication," in 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), 2024, pp. 1–6.

[14] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), 2019, pp. 3645–3650.

[15] L. Lannelongue, J. Grealey, and M. Inouye, "Green algorithms: Quantifying the carbon footprint of computation," Advanced Science, vol. 8, no. 12, p. 2100707, 2021.

[16] United Nations Institute for Training and Research (UNITAR), "Global e-waste monitor 2024," https://www.unitar.org/about/news-stories/press/global-e-waste-monitor-2024, 2024, accessed 5 July 2025.

[17] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19), 2019, pp. 220–229.