MASALIQ
Jurnal Pendidikan dan Sains

e-ISSN: 2808-8115
p-ISSN: 2809-1051

Terindeks: Sinta 5, Copernicus, Dimensions, Scilit, Lens, Crossref, Garuda, Google Scholar, Base, etc

https://doi.org/10.58578/masaliq.v6i1.8880

OPERATIONALIZING AI ETHICS: A THEMATIC SYNTHESIS FOR ADAPTIVE AND LAYERED GOVERNANCE

Irhas
Universitas Bumigora, Mataram, Indonesia
irhas@universitasbumigora.ac.id

Article Info:

Submitted:	Revised:	Accepted:	Published:
Dec 11, 2025	Jan 3, 2026	Jan 15, 2026	Jan 20, 2026

Abstract

This study is motivated by the growing discourse on artificial intelligence (AI) ethics that has not yet been fully operationalized in practice, particularly in the health and education sectors in the era of generative AI. It aims to synthesize the literature on the operationalization of AI ethics in order to propose a new governance model that is more adaptive and context-sensitive. A structured thematic synthesis was conducted on peer-reviewed studies published between 2020 and 2025. The analysis reveals a central paradox: a strong global consensus on the principles of justice, transparency, accountability, and privacy stands in stark contrast to significant gaps in their implementation. The emerging challenges are highly context-dependent: in the health sector, they center on technical safety and relational accountability, whereas in the education sector they are more closely related to epistemic integrity and equitable access. The advent of generative AI exacerbates these issues by introducing new systemic risks, such as large-scale hallucinations and an increasingly profound crisis of clarity. In response to the limitations of static compliance-based approaches, this study proposes the Adaptive and Layered AI Governance Model (ALAIGoM), a model of adaptive and layered AI governance that integrates operationalized ethical principles, multi-level governance structures, and a continuous ethics life

Masaliq Journal is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License

cycle to enable context-sensitive, adaptive oversight. The model is grounded in proactive trust-building, human-centered design, and mechanisms that ensure acceptability. Overall, this study offers a coherent roadmap for reconfiguring AI ethics from a mere normative checklist into an embedded, adaptive, and responsive governance process.

Keywords: Artificial Intelligence Ethics; Structured Thematic Synthesis; Adaptive AI Governance Model; Health Sector; Education Sector

INTRODUCTION

The development of Artificial Intelligence (AI) has reached a transformative phase where the technology no longer functions merely as a computational aid, but has become an architectural force actively reshaping the socio-technical foundations of contemporary society. Its infiltration into critical domains such as healthcare, education, judicial systems, and labor market mechanisms has reached a depth that alters decision-making paradigms, reconfigures professional relationships, and redistributes agency (Floridi & Cowls, 2021; Kaplan & Haenlein, 2020). The dominant narrative often emphasizes its utopian promise: diagnostic algorithms with superhuman acuity, personally adaptive learning systems, and unprecedented economic efficiency (Abràmoff et al., 2022; Davenport et al., 2020). However, behind this spectacular technical progress lies a complex and tension-filled ethical landscape, where the same innovation acts as a catalyst for systemic risks to individual autonomy, procedural justice, privacy, and the integrity of social institutions (Fukuda-Parr & Gibbons, 2021; Stahl et al., 2022). This dynamic reveals a fundamental paradox wherein technological advancement does not progress linearly with ethical progress, thus requiring reflective navigation through a terrain of often conflicting values.

The emergence of generative AI and Large Language Models (LLMs) has served as both an accelerator and a disruptor to existing AI ethics discourse. This technology blurs traditional boundaries between human and machine production, challenging established conceptions of creativity, authenticity, and expertise (Dowling & Lucey, 2023; Sison et al., 2024). Its capacity to generate convincing text, code, and imagery raises urgent normative questions that touch the core of social institutions: How can concepts of authorship and intellectual property be upheld? How can academic integrity and epistemic foundations be maintained amidst easily accessible generative capabilities? Who is accountable for the mass

production of misinformation and hallucinations? (Harrer, 2023; Wach et al., 2023). The challenges posed by generative AI do not merely represent an escalation of previous computational ethics issues, but rather constitute a qualitative shift demanding a reconfiguration of conceptual and governance frameworks.

In response to this complexity, a proliferation of principles, guidelines, and ethical frameworks from various jurisdictions and organizations has occurred (Jobin et al., 2019). However, academic and policy discourse often remains fragmented within disciplinary silos. Philosophical technology discourse tends to focus on abstract principles like accountability and algorithmic justice (Ashok et al., 2022), while practitioners in fields like healthcare grapple with diagnostic bias and doctor-patient relationship dynamics (Amann et al., 2020; Aquino et al., 2023). In the realm of education, debate concentrates on digital literacy, access disparities, and threats to academic honesty (Chiu et al., 2024; Vetter et al., 2024). This fragmentation yields isolated approaches, where solutions designed for one context may overlook or even exacerbate problems in another, simultaneously hindering the development of a holistic understanding of the systemic patterns that require robust and adaptable governance principles.

This research bridges that gap through a comprehensive narrative review and thematic synthesis of recent multidisciplinary literature. Its contribution lies in integrating three analytical dimensions that are often examined in isolation: universal ethical principles, their context-specific manifestations in high-risk domains such as healthcare and education, and the disruptive ethical challenges introduced by generative AI. This tripartite perspective enables a critical examination of how core principles such as transparency and justice are stressed, reshaped, and contested when applied within clinical environments, educational settings, and the generative capabilities of large language models. Such integrative synthesis remains relatively scarce in existing scholarship and highlights the need to move beyond fragmented, domain-specific analyses. Consequently, the findings underscore the limitations of static, checklist-based ethical frameworks and point toward the necessity of more adaptive, context-sensitive, and lifecycle-oriented approaches to AI governance (de Almeida et al., 2021; Morley et al., 2021).

Despite the growing body of literature on AI ethics and governance, existing studies remain fragmented across disciplines and sectors. Previous research has tended to focus either on normative ethical principles or on isolated implementation challenges within

domains such as healthcare and education, without integrating these perspectives or fully addressing the disruptive implications of generative AI. Consequently, a significant gap persists in understanding how universal ethical principles can be operationalized across high-risk sectors while remaining responsive to context-specific demands and the systemic risks introduced by large language models. Addressing this gap, this study synthesizes core AI ethical principles, their sector-specific manifestations in healthcare and education, and the ethical disruptions posed by generative AI. Grounded in AI ethics theory, human-centered AI, and adaptive governance perspectives, this research aims to develop an Adaptive and Layered AI Governance Model (ALAIGoM) to support context-sensitive and lifecycle-based ethical oversight in the generative AI era.

METHOD

This study employs a qualitative research approach based on a literature review methodology. Designed as a literature-based qualitative inquiry, it aims to synthesize conceptual, ethical, and governance-related insights concerning artificial intelligence in the generative AI era. This approach is appropriate for examining complex normative and contextual issues that cannot be adequately captured through quantitative measurement alone (Creswell & Poth, 2016).

The research design adopts a narrative review integrated with thematic synthesis. This design enables the integration of multidisciplinary scholarly perspectives and facilitates a systematic interpretation of ethical principles, governance frameworks, and sector-specific challenges in healthcare and education. A narrative review is particularly suitable for mapping conceptual developments across heterogeneous bodies of literature, while thematic synthesis supports theory-informed interpretation and pattern identification across studies (Braun & Clarke, 2006; Popay et al., 2006).

The units of analysis in this study consist of peer-reviewed journal articles rather than human participants. A purposive selection technique was applied to identify relevant literature published between 2020 and 2025, following established qualitative sampling principles for conceptual review studies (Patton, 2015). This timeframe was deliberately chosen to capture the evolution of AI ethics discussions following the emergence of generative models.

Data were collected through a structured literature selection protocol, which included predefined inclusion and exclusion criteria, keyword-based search strategies, and a thematic coding framework. The literature screening process involved title, abstract, and full-text review stages to ensure analytical relevance and methodological transparency (Snyder, 2019). The resulting corpus represents three primary analytical axes: global governance principles, implementation in health and education sectors, and challenges posed by generative AI.

The selected literature was analyzed using inductive thematic synthesis. The analysis involved initial familiarization with the texts, open coding, and the development of higher-order themes through iterative comparison. This analytical procedure follows established thematic analysis principles widely used in qualitative synthesis research (Braun & Clarke, 2006). The entire process was reflexive and supported by researcher triangulation and documented analytical protocols to ensure credibility.

RESULTS

Based on a systematic thematic synthesis of recent multidisciplinary literature, this study identifies a complex and multi-layered landscape of AI ethics characterized by recurring patterns across domains and technological contexts. The analysis reveals three primary thematic clusters: consensus on core ethical principles alongside persistent implementation gaps, sector-specific manifestations of ethical challenges in healthcare and education, and the expanding ethical complexity introduced by generative AI.

1. Consensus on Ethical Principles and Implementation Challenges

The literature demonstrates a broad convergence around foundational ethical principles, including transparency, fairness, accountability, privacy, and safety, which are consistently identified as central to responsible AI development (Floridi & Cowls, 2021; Jobin et al., 2019). Across studies, these principles appear frequently in policy frameworks, ethical guidelines, and governance initiatives. However, the synthesis also shows that their operationalization remains uneven and context-dependent.

Transparency and explainability are commonly reported as prerequisites for trust and oversight. In healthcare contexts, studies emphasize the need for explanations that are interpretable by clinicians to support AI-assisted decision-making (Amann et al., 2020). At the same time, several analyses note a growing tension between model complexity and

interpretability, particularly as advanced AI systems demonstrate increased predictive performance alongside reduced explainability (Gevaert, 2022). Similar implementation challenges are reported for fairness, where multiple studies document how historical and structural biases embedded in training data may be reproduced or amplified by algorithmic systems, including in recruitment and screening applications (Chen, 2023; Starke et al., 2022).

The issue of accountability is frequently described as fragmented. The reviewed literature highlights persistent ambiguity regarding responsibility for AI-driven decisions, particularly in cases involving harm or error, with liability often distributed across developers, platform providers, and end-users (Tóth et al., 2022). In parallel, privacy concerns are consistently raised in relation to large-scale data collection and processing practices, which complicate traditional models of informed consent and individual data control (de Almeida et al., 2021; Landers & Behrend, 2023; Morley et al., 2021).

Several governance-oriented studies document the emergence of mechanisms such as ethical guidelines, algorithmic auditing, and impact assessment frameworks as responses to these challenges (de Almeida et al., 2021; Landers & Behrend, 2023; Morley et al., 2021). Nonetheless, the literature reports variability in how such mechanisms are implemented and enforced across institutional and regulatory contexts (Bleher & Braun, 2023).

2. Sector-Specific Ethical Challenges in Healthcare and Education

The synthesis further indicates that ethical principles acquire distinct meanings and operational challenges when applied within specific sectors. In healthcare, the dominant concerns relate to patient safety, equity, and clinical accountability. Multiple studies report that AI systems trained on non-representative datasets may exhibit reduced accuracy for certain demographic groups, raising concerns about unequal treatment outcomes (Elendu et al., 2023; Mennella et al., 2024). Questions of responsibility also become salient when AI-supported clinical decisions contribute to adverse outcomes, with legal accountability often remaining unclear (Cobianchi et al., 2022). Privacy issues are particularly pronounced due to the sensitivity of medical data and the scale of data aggregation required for model development (Larson et al., 2020).

In educational contexts, the literature emphasizes challenges related to access, integrity, and pedagogical transformation. Several studies highlight disparities in infrastructure and digital literacy that influence the equitable adoption of AI-supported learning technologies (Zembylas, 2023). The rapid diffusion of generative AI tools is also

associated with renewed concerns about academic integrity, as automated content generation complicates traditional understandings of authorship and originality (Vetter et al., 2024). Additionally, studies note the growing importance of AI literacy among educators and learners to support informed and critical engagement with AI systems (Chiu et al., 2024; Kong et al., 2024).

A comparative overview of these sector-specific patterns is presented in Table 1, which summarizes how key ethical principles manifest differently across healthcare and education, while also highlighting emerging dimensions associated with generative AI.

Table 1. Sector-specific manifestations of key AI ethical challenges in healthcare and education, including generative AI–related complexities

Ethical Principle	Challenge in Healthcare	Challenge in Education	New Dimensions & Added Complexity*
Fairness & Non-Discrimination	Diagnostic bias against minority demographic groups (Elendu et al., 2023)	Gaps in access and literacy that exacerbate inequality (Zembylas, 2023).	Amplification and new scales of bias due to training on massive, uncurated datasets (Harrer, 2023).
Accountability	Blurring of legal liability among developers, institutions, and clinicians (Cobianchi et al., 2022).	Difficulty in establishing liability for AI-assisted academic integrity breaches (Vetter et al., 2024).	Further obfuscated accountability landscape due to difficulties in tracing sources of hallucinated or harmful content (Sison et al., 2024).
Transparency & Explainability	Need for explainability comprehensible to clinicians for adoption (Amann et al., 2020).	Necessity of AI literacy as a foundation for critical understanding (Chiu et al., 2024).	Deepened explainability crisis caused by the "black box" complexity of LLMs (Sison et al., 2024).
Privacy	Vulnerability of highly sensitive patient data in training datasets (Larson et al., 2020).	Protection of educational data and student learning profiles (Lameras & Arnab, 2021).	New collective privacy dilemmas arising from massive data scraping practices for model training (Dowling & Lucey, 2023).

3. The Disruptive Dimensions of Generative AI

The reviewed literature consistently identifies generative AI, particularly large language models (LLMs), as a significant source of additional ethical complexity that extends beyond previously documented AI governance challenges. Rather than introducing entirely separate concerns, generative AI is reported to intensify existing ethical tensions while simultaneously generating novel risk configurations ((Harrer, 2023; Sison et al., 2024)

One prominent pattern concerns large-scale content generation. Multiple studies document the capacity of LLMs to produce text, code, and visual materials at high volume and apparent coherence, which has been associated with increased risks of misinformation, fabrication, and the dissemination of factually inaccurate yet plausible outputs, commonly referred to as hallucinations (Harrer, 2023). These characteristics are reported across both healthcare- and education-related contexts, where the credibility of information is particularly consequential.

The literature also highlights a deepening explainability challenge. Compared to earlier AI systems, LLM architectures are frequently described as exhibiting heightened opacity due to their scale, emergent behaviors, and non-linear learning processes (Sison et al., 2024). Studies report that this opacity complicates efforts to trace decision pathways, assess system reliability, and provide meaningful explanations to end-users, especially in high-risk applications.

Data governance emerges as another recurring concern. Several analyses note that the training of generative AI models commonly relies on large-scale data scraping practices, raising unresolved questions regarding consent, data ownership, and intellectual property rights (Dowling & Lucey, 2023). These practices are reported to challenge existing regulatory frameworks, particularly those grounded in individual consent and data minimization principles.

In addition, the literature documents broader socio-economic effects associated with generative AI deployment. Studies indicate potential disruptions to professional roles, creative labor, and skill valuation, with implications for both healthcare work practices and educational objectives (McLean et al., 2023). These developments are described as contributing to a more dynamic and uncertain ethical landscape compared to earlier generations of AI systems.

4. Cross-Sectoral Ethical Foundations Identified in the Literature

Across domains and technological contexts, the synthesis identifies several recurring ethical orientations that appear consistently in the reviewed studies. These orientations are not presented as sector-specific solutions but as cross-cutting patterns observed in discussions of responsible AI deployment.

First, trust is frequently referenced as a central condition for AI acceptance and sustained use. The literature describes trust as associated with system reliability, transparency of system limitations, and consistency between system behavior and stated ethical commitments (Choung et al., 2023; Li et al., 2024). Studies across healthcare and education report that perceived deficits in any of these dimensions are linked to resistance, misuse, or disengagement from AI-supported systems.

Second, a human-centered orientation is widely documented. Research grounded in Human-Centered AI (HCAI) emphasizes the positioning of AI systems as supportive tools rather than autonomous replacements for human judgment (Bingley et al., 2023). Across sectors, studies report increased attention to participatory design approaches and the involvement of domain professionals, such as clinicians and educators, in system development and evaluation processes.

Third, mechanisms of contestability are repeatedly identified as relevant to ethical oversight. The literature notes that the availability of procedures allowing affected individuals to question, challenge, or seek redress for AI-assisted decisions is associated with perceptions of fairness and accountability (Lyons et al., 2021). These mechanisms are discussed in relation to both institutional governance arrangements and system-level design features.

5. Overall Synthesis of Findings

Taken together, the results indicate that the contemporary AI ethics landscape is characterized by a recurring pattern: broad normative convergence at the level of ethical principles coexists with substantial variability in implementation across sectors and technologies. While principles such as transparency, fairness, accountability, and privacy are widely referenced and accepted (Floridi & Cowls, 2021), their operational realization differs markedly depending on institutional context, domain-specific risk profiles, and technological characteristics.

The synthesis further shows that healthcare and education exhibit distinct constellations of ethical challenges, shaped by their respective professional norms, data sensitivities, and accountability structures. At the same time, the emergence of generative AI introduces additional layers of complexity that cut across these domains, amplifying concerns related to explainability, data governance, and informational integrity.

Overall, the findings demonstrate that ethical challenges in AI are neither uniform nor static. Instead, they evolve through the interaction of universal principles, sectoral contexts, and rapidly advancing technological capabilities, as documented across the reviewed literature.

DISCUSSION

This thematic synthesis confirms a fundamental paradox in AI ethics: while there is strong normative consensus regarding core ethical principles, their implementation remains characterized by uncertainty and considerable variation across contexts. Building on this paradox, the discussion examines the implications of the findings by analyzing the persistent gap between normative declarations and operational practices, exploring governance requirements that are responsive to contextual and technological disruption, and articulating a conceptual framework derived from the synthesis.

1. Bridging the Principle–Implementation Gap: From Declaration to Institutionalization

The widespread agreement on foundational AI ethics principles represents an important normative milestone in guiding technological development (Jobin et al., 2019). However, as demonstrated by the results, this consensus does not automatically translate into effective practice. The central challenge lies less in the absence of ethical principles than in the limited institutional mechanisms available to embed these principles into enforceable standards, technical procedures, and accountability arrangements (Bleher & Braun, 2023).

In healthcare contexts, accountability extends beyond ethical guidance and is closely tied to unresolved issues of legal responsibility and risk allocation when algorithmic systems contribute to adverse outcomes (Cobianchi et al., 2022). In educational settings, fairness-related concerns are similarly shaped by structural conditions, particularly inequalities in access and digital literacy that may influence how AI technologies are adopted and experienced (Zembylas, 2023).

Taken together, the literature suggests a gradual shift in emphasis from the continued articulation of ethical principles toward their institutionalization within organizational and regulatory practices. Approaches such as Ethics as a Service illustrate one pathway by integrating ethical expertise and tools directly into development processes, thereby embedding ethical considerations throughout the technology lifecycle rather than treating them as a final compliance requirement (Morley et al., 2021). At the same time, the synthesis indicates that institutionalization efforts must remain sensitive to contextual variation across domains in order to be effective.

2. Designing Governance Responsive to Context and Technological Disruption

The results demonstrate that ethical challenges associated with AI differ substantially across application domains, supporting the view that governance approaches based solely on uniform regulatory instruments are likely to face limitations. Governance frameworks that rely exclusively on centralized, top-down regulation may struggle to accommodate the operational realities of sectors such as healthcare and education (de Almeida et al., 2021).

Instead, the literature points toward multi-layered governance arrangements that balance overarching normative coherence with sector-specific adaptability (Stahl et al., 2022). At a macro level, regulatory instruments establish baseline protections, safety requirements, and fundamental rights. At meso and micro levels, sectoral and organizational actors develop operational guidelines, audit mechanisms, and ethical protocols that reflect domain-specific risks and professional norms, such as clinical governance models in healthcare or academic integrity frameworks in education (Marcus et al., 2024; Vetter et al., 2024).

The rapid evolution of generative AI further reinforces the importance of governance approaches that incorporate adaptive and anticipatory capacities. Studies emphasize the role of ex ante mechanisms, including regulatory sandboxes, ethical and fundamental rights impact assessments, and continuous post-deployment monitoring, as ways to address risks before they materialize at scale (Landers & Behrend, 2023). In the case of LLM-generated misinformation and harmful content, collaborative arrangements involving regulators, platforms, and researchers are frequently identified as necessary for effective mitigation (Harrer, 2023). These findings suggest that responsiveness to technological disruption depends not only on regulatory design but also on sustained institutional coordination.

3. Integrating Cross-Sectoral Foundations: Trust, Human-Centered Design, and Contestability

Across the reviewed literature, three cross-sectoral foundations consistently emerge as central to ethical AI governance: trust, human-centered design, and contestability. Rather than functioning as supplementary considerations, these elements appear as structural conditions that support the legitimacy and effectiveness of governance arrangements.

Trust is widely described as a prerequisite for the adoption and sustained use of AI systems. Studies associate trust with system reliability, transparency regarding capabilities and limitations, and observable alignment between system behavior and declared ethical commitments (Li et al., 2024). In the absence of these conditions, resistance and disengagement from AI-supported systems are commonly reported across both healthcare and education contexts.

Human-centered design principles further emphasize the role of AI as a supportive technology that enhances, rather than replaces, human judgment and agency. Research grounded in Human-Centered AI highlights the importance of participatory design processes that involve end-users and affected stakeholders throughout system development and evaluation (Bingley et al., 2023). In practical terms, this orientation is reflected in educational tools designed to support teachers' pedagogical autonomy and clinical systems intended to augment professional decision-making rather than automate it (Lameras & Arnab, 2021).

Contestability represents a third foundational element, closely linked to perceptions of fairness and accountability. The literature underscores the importance of accessible mechanisms that allow individuals to question, challenge, and seek redress for AI-assisted decisions (Lyons et al., 2021). Together, these three foundations are frequently described as mutually reinforcing, with human-centered design supporting trust, and contestability mechanisms providing procedural safeguards when trust is challenged.

4. Conceptual Implication: An Adaptive Layered AI Governance Model (ALAIGoM)

As a conceptual implication of the synthesis, this study outlines an Adaptive Layered AI Governance Model (ALAIGoM) that integrates the patterns identified across the reviewed literature. The model addresses the tension between universal ethical principles and contextual variability by organizing governance along three interrelated dimensions.

The content dimension concerns the operationalization of ethical principles through measurable criteria, technical standards, and audit mechanisms. For example, fairness

requires explicit definitions of bias and standardized reporting metrics, while transparency is supported through documentation practices and context-appropriate explainability tools (Amann et al., 2020; Landers & Behrend, 2023).

The structural dimension reflects a layered governance architecture encompassing global norms, sector-specific guidelines, and organizational implementation practices. This layered arrangement allows alignment with shared ethical commitments while accommodating the distinct risk profiles and institutional logics of different domains (de Almeida et al., 2021).

The process dimension emphasizes the embedding of ethical evaluation throughout the technology lifecycle, from early design to deployment and decommissioning. This includes impact assessments, continuous monitoring, and mechanisms for revision and correction as systems evolve (Morley et al., 2021). The literature consistently highlights the importance of multi-stakeholder engagement in sustaining such processes over time (Kusters et al., 2020).

5. Implications in the Context of Generative AI

The findings indicate that generative AI intensifies existing ethical challenges while introducing new forms of systemic risk, including large-scale hallucinations, heightened opacity, and unresolved intellectual property concerns. In this context, governance approaches that remain static or primarily reactive appear increasingly misaligned with the pace and complexity of technological change.

By emphasizing adaptability, contextual sensitivity, and continuous learning, the ALAIGoM framework reflects the governance orientations most frequently advocated in recent scholarship. Its primary contribution lies in synthesizing these orientations into an integrated perspective that connects ethical principles, institutional structures, and lifecycle processes. In doing so, the discussion underscores the importance of governance arrangements that support not only technical innovation but also social trust, accountability, and human well-being in the evolving generative AI landscape.

6. Analytical Boundaries of the Study

This thematic synthesis is subject to several analytical boundaries that should be considered when interpreting the findings. First, the analysis is based exclusively on peer-reviewed literature published between 2020 and 2025, which may limit sensitivity to very

recent regulatory developments, industry practices, or grey literature not yet captured in academic discourse.

Second, as a narrative and thematic synthesis, this study does not empirically test the proposed Adaptive Layered AI Governance Model (ALAIGoM). Consequently, the framework should be understood as a conceptual and integrative contribution rather than an empirically validated governance solution.

Finally, the focus on healthcare and education, while analytically justified due to their high-risk characteristics, constrains the immediate transferability of the findings to other domains such as finance, security, or public administration. Future empirical research is therefore required to examine how the proposed framework performs across different institutional and regulatory contexts.

CONCLUSION

This thematic synthesis demonstrates that contemporary AI ethics is shaped by a persistent tension between strong normative consensus and deeply context-dependent implementation challenges. While core principles such as transparency, fairness, accountability, and privacy are widely endorsed across the literature, their operationalization varies significantly across high-risk domains such as healthcare and education. The findings further show that generative AI intensifies existing ethical concerns and introduces new systemic risks, including large-scale hallucinations, explainability limitations, and unresolved data governance and intellectual property issues.

This study contributes to the field of AI ethics and governance in three main ways. First, it offers an integrated synthesis that connects universal ethical principles, sector-specific manifestations, and the disruptive implications of generative AI, addressing a fragmentation gap in prior literature. Second, it advances a conceptual governance contribution through the Adaptive Layered AI Governance Model (ALAIGoM), which translates ethical principles into operational content, contextual governance structures, and a continuous ethics lifecycle. Third, it foregrounds trust, human-centered design, and contestability as cross-sectoral foundations that are essential for legitimate and effective AI governance.

Based on these contributions, future research is encouraged to empirically examine the applicability of the proposed model through sector-specific case studies, to develop standardized metrics for ethics auditing and lifecycle monitoring, and to conduct comparative analyses of AI governance practices across regulatory contexts. Such efforts are essential to ensuring that AI governance remains responsive to rapid technological change while remaining aligned with fundamental human values.

REFERENCES

Abràmoff, M. D., Roehrenbeck, C., Trujillo, S., Goldstein, J., Graves, A. S., Repka, M. X., & Silva, E. Z., III. (2022). A reimbursement framework for artificial intelligence in healthcare. *npj Digital Medicine, 5*(1), Article 72. https://doi.org/10.1038/s41746-022-00621-w

Amann, J., Blasimme, A., Vayena, E., Frey, D., Madai, V. I., & Precise4Q Consortium. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making, 20*(1), 310. https://doi.org/10.1186/s12911-020-01332-6

Aquino, Y. S. J., Rogers, W. A., Braunack-Mayer, A., Frazer, H., Win, K. T., Houssami, N., Degeling, C., Semsarian, C., & Carter, S. M. (2023). Utopia versus dystopia: Professional perspectives on the impact of healthcare artificial intelligence on clinical roles and skills. *International Journal of Medical Informatics, 169*, 104903. https://doi.org/10.1016/j.ijmedinf.2022.104903

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for artificial intelligence and digital technologies. *International Journal of Information Management, 62*, 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Bingley, W. J., Curtis, C., Lockey, S., Bialkowski, A., Gillespie, N., Haslam, S. A., Ko, R. K. L., Steffens, N., Wiles, J., & Worthy, P. (2023). Where is the human in human-centered AI? Insights from developer priorities and user experiences. *Computers in Human Behavior, 141*, 107617. https://doi.org/10.1016/j.chb.2022.107617

Bleher, H., & Braun, M. (2023). Reflections on putting AI ethics into practice: How three AI ethics approaches conceptualize theory and practice. *Science and Engineering Ethics, 29*(3), 21. https://doi.org/10.1007/s11948-023-00443-3

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Chen, Z. (2023). Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications, 10*(1), 1–12. https://doi.org/10.1057/s41599-023-02079-x

Chiu, T. K. F., Ahmad, Z., Ismailov, M., & Sanusi, I. T. (2024). What are artificial intelligence literacy and competency? A comprehensive framework to support them. *Computers and Education Open, 6*, 100171. https://doi.org/10.1016/j.caeo.2024.100171

Choung, H., David, P., & Ross, A. (2023). Trust and ethics in AI. *AI & Society, 38*(2), 733–745. https://doi.org/10.1007/s00146-022-01473-4

Cobianchi, L., Verde, J. M., Loftus, T. J., Piccolo, D., Dal Mas, F., Mascagni, P., Garcia Vazquez, A., Ansaloni, L., Marseglia, G. R., Massaro, M., Gallix, B., Padoy, N., Peter, A., & Kaafarani, H. M. (2022). Artificial intelligence and surgery: Ethical dilemmas and open issues. *Journal of the American College of Surgeons, 235*(2), 268–275. https://doi.org/10.1097/XCS.0000000000000242

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). Sage Publications.

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science, 48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

de Almeida, P. G. R., dos Santos, C. D., & Farias, J. S. (2021). Artificial intelligence regulation: A framework for governance. *Ethics and Information Technology, 23*(3), 505–525. https://doi.org/10.1007/s10676-021-09593-z

Dowling, M., & Lucey, B. (2023). ChatGPT for (finance) research: The Bananarama conjecture. *Finance Research Letters, 53*, 103662. https://doi.org/10.1016/j.frl.2023.103662

Elendu, C., Amaechi, D. C., Elendu, T. C., Jingwa, K. A., Okoye, O. K., John Okah, M., Ladele, J. A., Farah, A. H., & Alimi, H. A. (2023). Ethical implications of AI and robotics in healthcare: A review. *Medicine, 102*(50), e36671. https://doi.org/10.1097/MD.0000000000036671

Floridi, L., & Cowls, J. (2021). A unified framework of five principles for AI in society. In *Philosophical Studies Series* (pp. 5–17). Springer International Publishing. https://doi.org/10.1007/978-3-030-81907-1_2

Fukuda-Parr, S., & Gibbons, E. (2021). Emerging consensus on "ethical AI": Human rights critique of stakeholder guidelines. *Global Policy, 12*(S6), 32–44. https://doi.org/10.1111/1758-5899.12965

Gevaert, C. M. (2022). Explainable AI for earth observation: A review including societal and regulatory perspectives. *International Journal of Applied Earth Observation and Geoinformation, 112*, 102869. https://doi.org/10.1016/j.jag.2022.102869

Harrer, S. (2023). Attention is not all you need: The complicated case of ethically using large language models in healthcare and medicine. *EBioMedicine, 90*, 104512. https://doi.org/10.1016/j.ebiom.2023.104512

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

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons, 63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Kong, S.-C., Cheung, M.-Y. W., & Tsang, O. (2024). Developing an artificial intelligence literacy framework: Evaluation of a literacy course for senior secondary students using a project-based learning approach. *Computers and Education: Artificial Intelligence, 6*, 100214. https://doi.org/10.1016/j.caeai.2024.100214

Kusters, R., Misevic, D., Berry, H., Cully, A., Le Cunff, Y., Dandoy, L., Díaz-Rodríguez, N., Ficher, M., Grizou, J., Othmani, A., Palpanas, T., Komorowski, M., Loiseau, P., Moulin-Frier, C., Nanini, S., Quercia, D., Sebag, M., Soulié-Fogelman, F., Taleb, S.,

... Wehbi, F. (2020). Interdisciplinary research in artificial intelligence: Challenges and opportunities. *Frontiers in Big Data, 3*, 577974. https://doi.org/10.3389/fdata.2020.577974

Lameras, P., & Arnab, S. (2021). Power to the teachers: An exploratory review on artificial intelligence in education. *Information, 13*(1), 14. https://doi.org/10.3390/info13010014

Landers, R. N., & Behrend, T. S. (2023). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist, 78*(1), 36–49. https://doi.org/10.1037/amp0000972

Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., & Langlotz, C. P. (2020). Ethics of using and sharing clinical imaging data for artificial intelligence: A proposed framework. *Radiology, 295*(3), 675–682. https://doi.org/10.1148/radiol.2020192536

Li, Y., Wu, B., Huang, Y., & Luan, S. (2024). Developing trustworthy artificial intelligence: Insights from research on interpersonal, human-automation, and human-AI trust. *Frontiers in Psychology, 15*, 1382693. https://doi.org/10.3389/fpsyg.2024.1382693

Lyons, H., Velloso, E., & Miller, T. (2021). Conceptualising contestability. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–25. https://doi.org/10.1145/3449180

Marcus, H. J., Ramirez, P. T., Khan, D. Z., Layard Horsfall, H., Hanrahan, J. G., Williams, S. C., Beard, D. J., Bhat, R., Catchpole, K., Cook, A., Hutchison, K., Martin, J., Melvin, T., Stoyanov, D., Rovers, M., Raison, N., Dasgupta, P., Noonan, D., Stocken, D., ... IDEAL Robotics Colloquium. (2024). The IDEAL framework for surgical robotics: Development, comparative evaluation and long-term monitoring. *Nature Medicine, 30*(1), 61–75. https://doi.org/10.1038/s41591-023-02732-7

McLean, S., Read, G. J. M., Thompson, J., Baber, C., Stanton, N. A., & Salmon, P. M. (2023). The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence, 35*(5), 649–663. https://doi.org/10.1080/0952813X.2021.1964003

Mennella, C., Maniscalco, U., De Pietro, G., & Esposito, M. (2024). Ethical and regulatory challenges of AI technologies in healthcare: A narrative review. *Heliyon, 10*(4), e26297. https://doi.org/10.1016/j.heliyon.2024.e26297

Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., & Floridi, L. (2021). Ethics as a Service: A pragmatic operationalisation of AI ethics. *Minds and Machines, 31*(2), 239–256. https://doi.org/10.1007/s11023-021-09563-w

Patton, M. Q. (2015). *Qualitative research and evaluation methods* (4th ed.). Sage Publications.

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). *Guidance on the conduct of narrative synthesis in systematic reviews: A product from the ESRC methods programme*. ESRC Methods Programme.

Sison, A. J. G., Daza, M. T., Gozalo-Brizuela, R., & Garrido-Merchán, E. C. (2024). ChatGPT: More than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) perspective. *International Journal of Human–Computer Interaction, 40*(17), 4853–4872. https://doi.org/10.1080/10447318.2023.2225931

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Stahl, B. C., Rodrigues, R., Santiago, N., & Macnish, K. (2022). A European Agency for Artificial Intelligence: Protecting fundamental rights and ethical values. *Computer Law & Security Review, 45*, 105661. https://doi.org/10.1016/j.clsr.2022.105661

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society, 9*(2), 20539517221115189. https://doi.org/10.1177/20539517221115189

Tóth, Z., Caruana, R., Gruber, T., & Loebbecke, C. (2022). The dawn of the AI robots: Towards a new framework of AI robot accountability. *Journal of Business Ethics, 178*(4), 895–916. https://doi.org/10.1007/s10551-022-05050-z

Vetter, M. A., Lucia, B., Jiang, J., & Othman, M. (2024). Towards a framework for local interrogation of AI ethics: A case study on text generators, academic integrity, and composing with ChatGPT. *Computers and Composition, 71*, 102831. https://doi.org/10.1016/j.compcom.2024.102831

Wach, K., Duong, C. D., Ejdys, J., Kazlauskaitė, R., Korzynski, P., Mazurek, G., Paliszkiewicz, J., & Ziemba, E. (2023). The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review, 11*(2), 7–30. https://doi.org/10.15678/EBER.2023.110201

Zembylas, M. (2023). A decolonial approach to AI in higher education teaching and learning: Strategies for undoing the ethics of digital neocolonialism. *Learning, Media and Technology, 48*(1), 25–37. https://doi.org/10.1080/17439884.2021.2010094