Beyond Bias: Teaching Ethical Reasoning and Governance Across the AI Lifecycle in the Information Systems Discipline*

Deepak Khazanchi[1,*,†], Anoop Mishra[1,†] and Vileru Tivexi[2,†]

[1] *University of Nebraska at Omaha, Omaha, Nebraska 68182, USA*
[2] *Doanne College, 1014 Boswell Avenue, Crete, Nebraska 68333, USA*

Abstract

As artificial intelligence (AI) is increasingly deployed in high-stakes domains such as healthcare, education, finance, and public administration, concerns related to fairness, transparency, trust, accountability, and governance have become central. Despite rapid technical advances, educational approaches have lagged in systematically integrating ethical reasoning into AI and data-driven curricula. This paper synthesizes recurring ethical dilemmas that emerge across the AI lifecycle and demonstrates how these dilemmas can serve as effective pedagogical entry points within Information Systems (IS) education. Drawing on scholarship in algorithmic fairness, trust in AI, ethical theory, and emerging AI governance frameworks, the paper examines persistent tensions such as fairness versus performance, privacy versus utility, autonomy versus control, transparency versus accuracy, and the dual-use nature of general-purpose AI through multiple ethical lenses. Building on this synthesis, the paper proposes evidence-based pedagogical strategies, including case-based, project-based, interdisciplinary, and experiential learning approaches, for embedding AI ethics directly into technical and business curricula. The paper argues that ethical reasoning, governance literacy, and technical competence must be developed in parallel to adequately prepare students for the responsible design, deployment, and governance of AI systems in practice.

Keywords

AI ethics, AI fairness, algorithmic fairness, pedagogy, teaching, IS education

1. Introduction

Artificial intelligence (AI) and machine learning (ML) systems are now embedded in a wide range of organizational and societal decision-making contexts, including healthcare diagnosis, hiring, lending, education, and public governance. While these systems promise gains in efficiency, scalability, and analytical insight, empirical evidence increasingly demonstrates that AI systems can reproduce and amplify historical inequities when trained on biased data or

NWISEd 2026: Workshop on Co-Creating New Ways of Information Systems Education, July 1-3, 2026, Swansea, United Kingdom

[1*] Corresponding author.
[†] These authors contributed equally.
khazanchi@unomaha.edu (D. Khazanchi); amishra@unomaha.edu (A. Mishra); vileru.tivexi@doane.edu (V. Tivexi)
0000-0002-2675-2871 (D. Khazanchi); 0000-0002-8264-6256 (A. Mishra)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-4333/paper4.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-19

deployed without sufficient oversight [33]. These challenges extend beyond technical limitations to questions of trust, accountability, and legitimacy.

This paper extends and substantially deepens ideas previously introduced in an invited keynote at the Third Workshop on Co-Creating New Ways of Information Systems Education (NWISEd 2025) [1]. The keynote argued that moving "beyond bias" requires equipping future professionals with ethical reasoning capabilities, fairness awareness, and governance literacy alongside technical competence. The present paper expands that contribution by integrating research on algorithmic fairness, public trust in AI, and pedagogy, and by articulating how these themes can be embedded systematically in information systems education.

The core argument advanced here is that AI ethics cannot be treated as a peripheral topic or compliance exercise. Instead, ethical reasoning, fairness assessment, and governance mechanisms must be embedded across the AI lifecycle and throughout educational curricula. Addressing AI ethics in this holistic manner is essential for preparing students to design, deploy, evaluate, and govern AI systems responsibly in practice.

2. Ethical Challenges of Contemporary AI

AI systems raise a range of ethical challenges that manifest across the lifecycle of data collection, model development, deployment, and use. Among the most prominent challenges are biased and incomplete data, lack of interpretability and transparency, privacy erosion, surveillance, security vulnerabilities, and unintended consequences. In healthcare, predictive algorithms have been shown to disadvantage marginalized populations because historical healthcare spending and access patterns were used as proxies for clinical need [23].

Beyond bias, generative AI systems introduce additional risks related to hallucination, misinformation, and misplaced trust. Empirical studies and practitioner reports indicate that large language models are more likely to hallucinate when responding to complex reasoning tasks, generating citations, or operating beyond their training data, raising significant concerns for educational and professional contexts [2]. Such risks are particularly problematic in high-stakes domains where users may over-rely on AI outputs.

Public trust in AI has therefore become a critical concern. Large-scale global surveys show that AI adoption continues to rise, yet trust in AI systems remains fragile and, in some regions, is declining due to concerns about misinformation, privacy, and bias [17]. Trust is also highly application-dependent, with healthcare AI viewed more favorably than AI used in human resource management or surveillance contexts [8]. These dynamics underscore the necessity of combining technical safeguards with ethical and governance education.

3. Core AI Ethical Dilemmas

AI ethics education benefits from framing recurring tensions as ethical dilemmas rather than as isolated technical defects.

3.1. Bias and Fairness

AI systems can produce discriminatory outcomes when training data or model design reflects existing social bias [33,35). A key dilemma is how to respond when a system is useful but still performs unevenly across demographic groups, especially in high-stakes settings where errors

can reinforce inequality [12,32]. For example, facial recognition systems have been shown to perform worse on darker-skinned individuals because of unbalanced training data and related design issues [12,19]. The central question is how to balance fairness, accountability, and business or technical constraints without causing harm to marginalized users [28].

3.2. Privacy versus Utility

Predictive accuracy often improves when AI systems use richer personal data, but that gain can create ethical tension between beneficence and individual rights. In education and healthcare, expanded data access can blur the line between useful insight and surveillance, especially when predictive analytics reveal sensitive inferences about people who never explicitly shared that information [2,21,26].

The key question is not whether AI can use more data, but whether the data use is necessary, proportionate, and transparent. In practice, the line is usually crossed when data collection or inference becomes more intrusive than the benefit justifies, or when people lose meaningful control over how their information is used [3,32].

3.3. Autonomy versus Control

AI systems raise difficult questions about how much decision-making power they should have in critical domains [14, 15]. A major dilemma is whether autonomous systems should follow strict human-defined rules or be allowed to optimize for the best outcome when trade-offs are unavoidable [6, 9]. For example, an autonomous vehicle facing a crash scenario may have to weigh the safety of passengers against the risk to a pedestrian, which raises questions about who is responsible for programming and justifying those choices [9, 13]. The central question is: should AI follow strict rules, or should it have enough autonomy to optimize for the best overall result?

3.4. Transparency versus Performance

High-performing models are often less interpretable, raising ethical concerns in domains where explanations are morally or legally required. Should AI systems prioritize explainability and transparency over raw accuracy? In high-stakes settings, the answer is usually not either/or: organizations should balance accuracy, explainability, and accountability based on context, because trustworthy AI requires tradeoffs to be handled transparently and justifiably.

For example, a black-box deep learning model that predicts loan approvals with 95% accuracy may still be problematic if it cannot explain why an applicant was rejected, since NIST notes that opaque systems can make risk measurement and accountability harder, especially in consequential uses [31,33]. The EU AI Act also emphasizes that high-risk AI should provide sufficient transparency so deployers can understand how it works and affected people can be informed of their rights [10]. If applicants demand explanations for loan denials, using a fully opaque model is ethically weak unless the institution can provide meaningful, case-specific explanations, oversight, and a defensible process for review. In practice, opacity is most acceptable only when the system is low-risk, tightly monitored, and not making or materially shaping high-impact decisions; for high-stakes AI, transparency and explainability should be treated as core requirements, not optional extras [32,33].

3.5. Dual-Use and Value Alignment

A particularly challenging class of ethical tensions in contemporary AI arises from dual-use dilemmas, where the same AI capability can generate substantial societal benefit while simultaneously enabling significant harm. General-purpose AI systems, especially large language models, exemplify this dilemma. Technologies developed to support legitimate applications such as education, research assistance, accessibility, and creative work can also be repurposed to produce scalable disinformation, deepfakes, automated propaganda, and targeted political manipulation at unprecedented speed and reach [31,25]. Empirical reviews show that generative AI plays a dual role: it accelerates the production and dissemination of false or misleading content while also offering tools for detection, verification, and media literacy, thereby intensifying both risks and responsibilities for developers and institutions [31].

These dual-use risks raise difficult questions about responsibility, disclosure, and governance in AI research and deployment. Restricting access to powerful AI systems may reduce opportunities for misuse, yet it can also limit beneficial innovation, exacerbate power asymmetries, and concentrate control in a small number of organizations. Conversely, broad or open release can promote transparency and democratization, but it may accelerate malicious use by well-resourced or coordinated actors. Recent work in research ethics emphasizes that AI dual-use risks must be considered early in the research lifecycle, with shared responsibility distributed among researchers, institutions, platforms, and oversight bodies rather than deferred to end-use mitigation alone [11]. This perspective aligns with emerging governance discussions that call for structured risk assessment, ethics review mechanisms, and anticipatory governance approaches tailored to AI's general-purpose nature.

Dual-use dilemmas are further compounded by value alignment challenges, as what constitutes "harmful use" varies considerably across cultural, political, and regulatory contexts. Global governance studies highlight that AI systems often encode dominant cultural norms and values, particularly those of Western, high-income societies, which can lead to misalignment when systems are deployed globally [43]. Empirical research on large language models demonstrates systematic cultural bias and uneven value alignment across regions, reinforcing concerns that ostensibly neutral AI systems may privilege certain worldviews while marginalizing others [39]. These findings underscore that value alignment is not a purely technical problem, but a socio-cultural and ethical one that requires continuous evaluation and engagement with diverse stakeholder communities.

Recent work on AI-driven disinformation and democratic resilience further demonstrates that dual-use risks are not hypothetical but already materially affecting public trust, political processes, and institutional legitimacy [37]. These dynamics raise a foundational ethical question: who gets to decide what counts as harmful AI use, and on what normative basis? Technical optimization alone cannot resolve this question. Instead, it requires ethical reasoning that integrates transparency, accountability, and representation, anticipates downstream societal impacts, and engages with evolving governance and regulatory frameworks [23]. Dual-use dilemmas therefore reinforce the necessity of embedding ethical reasoning and governance considerations alongside technical decision-making throughout the AI lifecycle.

4. Ethical Theories as Analytical Lenses

Ethical theories provide structured and complementary frameworks for analyzing AI dilemmas and for teaching systematic ethical reasoning in socio-technical contexts.

Deontological approaches, rooted in Kantian ethics, emphasize duties, rights, and moral constraints, implying that AI systems which violate fundamental human rights—such as privacy, autonomy, or non-discrimination—should not be deployed regardless of efficiency gains or aggregate benefits [22,15]. From this perspective, ethical acceptability is evaluated based on the nature of the action itself rather than its consequences.

In contrast, utilitarianism prioritizes outcomes and aggregate welfare, assessing AI systems based on their ability to maximize overall benefit or minimize harm. Utilitarian reasoning is frequently invoked in debates around AI deployment in areas such as healthcare allocation or autonomous vehicles, where tradeoffs may disadvantage specific groups if doing so increases total societal utility [29,10]. However, critics note that such approaches risk legitimizing systematic harm to minorities when efficiency dominates fairness considerations.

Virtue ethics shifts attention away from rules or outcomes toward the moral character, intentions, and practical wisdom of AI designers and decision makers. Drawing on Aristotelian foundations, virtue-based approaches emphasize qualities such as responsibility, honesty, fairness, and prudence in the design, development, and deployment of AI systems [4,23]. This perspective is particularly useful for examining professional conduct and organizational cultures shaping AI practice [23].

Rawlsian theories of justice emphasize the notion of fairness under conditions of uncertainty, asking whether AI-enabled decisions would be deemed acceptable if individuals did not know their own social position, demographic attributes, or access to resources. This "veil of ignorance" provides a powerful lens for evaluating algorithmic decision systems that exhibit disparate impact across groups [36,38]. Rawlsian reasoning has been influential in contemporary debates on algorithmic fairness and institutional legitimacy [38].

Finally, care ethics emphasizes relationships, empathy, context, and attentiveness to vulnerability, challenging overly abstract or impersonal approaches to AI governance. Care-based perspectives are especially salient in healthcare, education, and social services, where ethical evaluation must attend to relational harm, emotional impact, and contextual nuance rather than formal equality alone [18,41]. Care ethics provides an essential counterbalance to efficiency-driven AI adoption in human-centered domains [23].

In educational settings, systematically comparing how these ethical theories address the same AI dilemma helps learners recognize that ethical decision making in AI involves pluralism, tradeoffs, and reflective judgment, rather than definitive or purely technical answers. Such comparative analysis equips students to navigate real-world AI governance challenges that are value-laden, contested, and context-dependent.

5. Enhancing AI Ethics Education through Process Fairness

AI ethics education often emphasizes biased outcomes, that is, who benefits and who is disadvantaged by algorithmic decisions. While this focus on distributive fairness is essential, Rawlsian justice theory makes clear that ethical legitimacy also depends on *procedural fairness*: the fairness of the processes through which decisions are designed, explained, implemented,

and governed. the fairness of the processes through which decisions are designed, explained, implemented, and governed [29, 30]. The TAR framework, explicitly grounded in Rawlsian justice principles, integrates these two dimensions of fairness through its core constructs of Transparency, Accountability, and Representativeness [30]. In this framework, *transparency* is defined as "the degree to which the ML process information and functions are traceable, explainable, and has fidelity" [30]; *Accountability* is the "degree to which the ML process incorporates accuracy, correctability, consistency, ethicality, and has governance" [30]; and, *Representativeness* is defined as "the extent to which the ML development process utilizes training and testing data that reflects the diversity and distribution of the population to which it generalizes to and integrates the diversity of thought for algorithm engineering itself" [30].

From a *distributive fairness* perspective, the TAR framework directs attention to how AI systems allocate benefits, risks, and opportunities across individuals and groups under conditions of uncertainty. Drawing on Rawls's veil-of-ignorance reasoning, the TAR framework, as a fairness framework for the ML development process, encourages evaluators and students to ask whether an AI-enabled decision system would be regarded as fair without knowing one's own demographic position, expertise, power, or vulnerability [29, 30]. This framing provides a normative basis for examining disparate impact, group-level harms, and the limits of purely statistical fairness metrics when detached from institutional and societal context.

At the same time, the TAR framework focuses on procedural fairness, an aspect of ethical reasoning with AI that is frequently underdeveloped in computing education and practice. TAR defines fairness assessment as the function of transparency, accountability, and representativeness [29,30]. The Transparency dimension emphasizes whether AI decision rules, data practices, and limitations are sufficiently visible and intelligible to those affected, enabling justification and contestation. Accountability focuses on whether clear authority, oversight, and recourse mechanisms exist across the sociotechnical system.

Pedagogically, TAR enhances AI ethics education by shifting learning objectives from identifying biased outputs to evaluating justice-preserving decision processes. It offers instructors a coherent framework for integrating ethical theory, organizational governance, and technical design, helping students recognize that ethical AI failures often arise not from intent or optimization errors alone, but from unjust or weakly governed processes. By teaching both distributive and procedural fairness through the TAR framework [29,30], IS programs can better prepare students to navigate AI ethical dilemmas that cannot be resolved through technical fixes alone and require sustained institutional and moral judgment.

6. Pedagogical Strategies for Teaching AI Ethics

Effective AI ethics education increasingly emphasizes the integration of ethical reasoning within technical curricula rather than positioning ethics as a peripheral or abstract add-on. Recent scholarship underscores that students develop deeper ethical competence when ethical inquiry is embedded directly into the design, deployment, and evaluation of AI systems they are learning to build [30,45,44,1].

Case-based learning remains a central strategy, enabling students to engage with authentic sociotechnical dilemmas in domains such as hiring, healthcare, credit scoring, and generative AI–driven misinformation. Empirical studies demonstrate that well-designed AI ethics cases

help learners surface trade-offs among fairness, accountability, transparency, and performance, while also grappling with regulatory and organizational constraints [7]. This approach mirrors real-world decision contexts and supports transfer from classroom learning to professional practice [23].

Interdisciplinary framing has been shown to be especially critical, as AI ethics problems routinely span technical, legal, social, and philosophical dimensions. Integrative pedagogical models that draw on information systems, computer science, law, ethics, and the social sciences enable students to reason across stakeholder perspectives and institutional contexts [40,45]. Complementing this, project-based learning embeds ethical reflection directly into AI development tasks, such as dataset construction, model evaluation, and deployment decisions, thereby focusing on ethics as an intrinsic design constraint rather than an after-the-fact consideration [40].

Recent work also highlights the value of constructivist and experiential learning approaches, including simulations, role playing, and scaffolded stakeholder analysis. These approaches support students' progression from ethical awareness to applied ethical judgment by situating learning within realistic, socially embedded scenarios [24,20]. Stakeholder-based role play, in particular, has been shown to cultivate empathy, critical reflection, and an appreciation of power asymmetries in AI-enabled systems [45].

Collectively, this body of evidence suggests that AI ethics literacy must be developed in parallel with technical competence to prepare graduates for organizational and societal contexts in which AI systems are designed, governed, and contested. Educational strategies that combine technical rigor with experiential ethical engagement are more likely to produce practitioners capable of responsible AI decision making in practice [1,24,44].

7. Concluding Remarks

This paper has argued that teaching AI ethics in Information Systems must move decisively beyond a narrow focus on technical bias mitigation or post-hoc compliance. As AI systems increasingly mediate consequential decisions, ethical challenges emerge not only from data or model choices, but from how systems are designed, deployed, governed, and trusted within complex organizational and societal contexts. Preparing future professionals therefore requires embedding ethical reasoning, fairness awareness, and governance literacy alongside technical competence throughout the curriculum.

By framing AI ethics as a set of recurring ethical dilemmas, rather than isolated technical failures, the paper demonstrates how students can be guided to recognize and reason through tensions such as fairness versus performance, privacy versus utility, autonomy versus control, transparency versus accuracy, and the dual-use nature of general-purpose AI. The use of established ethical theories as analytical lenses further enables students to compare normative perspectives, appreciate tradeoffs, and develop reflective judgment rather than seeking definitive or purely technical answers.

The pedagogical strategies reviewed, case-based learning, project-based and interdisciplinary design tasks, and experiential approaches such as simulations and stakeholder role-play, illustrate practical ways to integrate ethics directly into AI-related coursework in IS and business programs. These approaches align ethical inquiry with the AI lifecycle itself, reinforcing the idea that ethical considerations are intrinsic design constraints rather than

external addons. Empirical evidence increasingly suggests that such integrated pedagogies are more effective in developing applied ethical competence and professional readiness.

Importantly, AI ethics education must also highlight governance and accountability structures. Risk-based regulatory frameworks, such as the EU AI Act, underscore that ethical AI is inseparable from institutional oversight, transparency obligations, and human responsibility [13]. Human-in-the-loop (HITL) mechanisms can support accountability and trust, but only when roles, authority, and responsibility are clearly defined rather than treated as symbolic safeguards. Incorporating governance concepts into education equips students to navigate regulatory expectations, organizational decision structures, and societal scrutiny surrounding AI deployment [2,23].

Taken together, this paper positions AI ethics education as a central responsibility of the Information Systems discipline. IS programs are uniquely situated to bridge technical design, organizational implementation, and governance concerns. By integrating ethical reasoning, technical understanding, and institutional context, IS education can play a pivotal role in shaping practitioners capable of designing and governing AI systems that are not only effective, but also fair, transparent, and socially legitimate. Moving beyond bias is therefore not simply an ethical aspiration, but an educational imperative.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly to perform a grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Adamakis, M., & Rachiotis, T. (2025). Artificial intelligence in higher education: A state-of-the-art overview of pedagogical integrity, artificial intelligence literacy, and policy integration. Encyclopedia, 5(4), Article 180. DOI: 10.3390/encyclopedia5040180

[2] Afroogh, Saleh, Akbari, Ali, Malone, Emmie, Kargar, Mohammadali & Alambeigi, Hananeh (2024). Trust in AI: Progress, Challenges, and Future Directions. Nature Humanities and Social Sciences Communications 11:1-30.

[3] American Psychological Association. (2025). Ethical guidance for AI in the professional practice of health service psychology. https://www.apa.org/topics/artificial-intelligence-machine-learning/ethical-guidance-ai-professional-practice

[4] Aristotle. Nicomachean Ethics. https://ia800506.us.archive.org/0/items/AristotlesNicomacheanEthics/Aristotles%20-%20Nicomachean%20Ethics.pdf

[5] Artificial Intelligence Act, Regulation (EU). (206). Article 50: Transparency Obligations for Providers and Deployers of Certain AI Systems. https://artificialintelligenceact.eu/article/50/I.

[6] Awad, E., Dsouza, S., Shariff, A., Rahwan, I., & Bonnefon, J.-F. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

[7] Barry, B. A., Sharp, R. R., & McGowan, M. L. (2025). Teaching the bioethics of information technologies and artificial intelligence in healthcare: Case-based learning for identifying

and addressing ethical issues. International Journal of Ethics Education, 10, 251–264. https://doi.org/10.1007/s40889-025-00218-3

[8] Benk, M., Kerstan, S., von Wangenheim, F., & Ferrario, A. (2025). Twenty-four years of empirical research on trust in AI. AI & Society, 40, 2083–2106.

[9] Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. Science 352,1573-1576. DOI:10.1126/science.aaf2654

[10] Brenneis, A. (2025). Assessing dual-use risks in AI research: necessity, challenges and mitigation strategies. Research Ethics, 21(2), 302–330.

[11] Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Researc*h, 81*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

[12] Ebers, M. (2025). Truly risk-based regulation of artificial intelligence: How to implement the EU AI Act. European Journal of Risk Regulation, 16(2), 684–703.

[13] European Commission. (2020). Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu

[14] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People -- An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines*, 28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5

[15] Gasser, U. and Almeida, V. A. F. (2017, November/December). "A Layered Model for AI Governance," in IEEE Internet Computing, vol. 21, no. 6, pp. 58-62, doi: 10.1109/MIC.2017.4180835.

[16] Gillespie, N., Lockey, S., Ward, T., Macdade, A., & Hassed, G. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. University of Melbourne & KPMG. https://kpmg.com/xx/en/our-insights/ai-and-technology/trust-attitudes-and-use-of-ai.html

[17] Gilligan, C. (1982). In a Different Voice. Harvard University Press.

[18] Harvard Kennedy School. (2020, November 3). Why racial bias is prevalent in facial recognition technology. The Case for Inclusive AI. https://jolt.law.harvard.edu/digest/why-racial-bias-is-prevalent-in-facial-recognition-technology

[19] Jinbao, Zhang. (2025). Integrating socio-scientific issues into AI ethics education: Curriculum framework and case study. US-China Education Review, 15(1), 1-12.

[20] Kairouz, P., et al. (2024). Predictive analytics and the collective dimensions of data protection. International Review of Law, Computers & Technology. https://doi.org/10.1080/17579961.2024.2313794

[21] Kant, I. (1785/2012). Groundwork of the Metaphysics of Morals. Translated by Mary Gregor and Jens Timmerman, Cambridge University Press. https://doi.org/10.1017/CBO9780511973741

[22] Khazanchi, D. (2025). Beyond Bias: Equipping the Next Generation to Navigate AI's Ethical Complexities. In Proceedings of the Third Workshop on Co-Creating New Ways of Information Systems Education (NWISEd 2025), CEUR Workshop Proceedings, Vol. 4076.

[23] Kim, S.-K., Kim, T.-Y., & Kim, K. (2025). Development and effectiveness verification of AI education datasets based on constructivist learning principles for enhancing AI literacy. Scientific Reports, 15, Article 10725. https://doi.org/10.1038/s41598-025-95802-4

[24] López-Borrull, A., & Lopezosa, C. (2025). Mapping the impact of generative AI on disinformation: Insights from a scoping review. Publications, 13(3), 33. https://doi.org/10.3390/publications13030033

[25] Mantelero, A. (2020). *Predictive privacy: Towards an applied ethics of data analytics. SSRN.* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3724185

[26] McDermid, John A., Jia, Yan, Porter, Zoe, Habli, Ibrahim (2021, October). Artificial intelligence explainability: the technical and ethical dimensions. Philos Trans A Math Phys Eng Sci 4: 379 (2207): 20200363. https://doi.org/10.1098/rsta.2020.0363.

[27] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, *54*(6), Article 115. https://doi.org/10.1145/3457607

[28] Mill, J. S. (1863). Utilitarianism. https://www.utilitarianism.com/jsmill-utilitarianism.pdf

[29] Mishra, A., Khazanchi, D., & Tripathi, A. (2026). Exploring demographic influences on fairness perceptions across the ML development lifecycle. AI and Ethics, 6(2), 201. DOI: 10.1007/s43681-026-01034-1.

[30] Mishra, A. (2025). Measuring perceived fairness of the machine learning (ML) development process: An empirical investigation. University of Nebraska at Omaha, USA. ProQuest Dissertations & Theses, 2025. 32169254.

[31] Montasari, R. (2024). The dual role of artificial intelligence in online disinformation: a critical analysis. In: Cyberspace, Cyberterrorism and the International Security in the Fourth Industrial Revolution: Threats, Assessment and Responses (pp. 229-240). Cham: Springer International Publishing.

[32] National Institute of Standards and Technology. (2023, January 25). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* https://doi.org/10.6028/NIST.AI.100-1.

[33] Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–453.

[34] Rafi, M.A., Shaboj, S.M., Miah, M.K., Rasul, I., Islam, M.R., & Ahmed, A. (2024). Explainable AI for Credit Risk Assessment: A Data-Driven Approach to Transparent Lending Decisions. Journal of Economics, Finance and Accounting Studies.

[35] Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, *169*(12), 866–872. https://doi.org/10.7326/M18-1990

[36] Rawls, J. (1971, 1999). A Theory of Justice. Harvard University Press.

[37] Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: Policy recommendations for democratic resilience. Frontiers in Artificial Intelligence, 8: 1569115. doi: 10.3389/frai.2025.1569115

[38] Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., & Vertesi, J. (2019, January). "Fairness and Abstraction in Sociotechnical Systems." FAT* 19: Proceedings of the Conference on Fairness, Accountability, and Transparency. Pages 59 – 68. https://doi.org/10.1145/3287560.3287598

[39] Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. PNAS Nexus, 3(9).

[40] Torres-Rivera, A. D., Rendón Peña, A. A., Díaz-Torres, S. T., & Díaz-Torres, L. A. (2025). Ethical integration of AI and STEAM pedagogies in higher education: A sustainable learning model for Society 5.0. Sustainability, 17(19), Article 8525. https://doi.org/10.3390/su17198525
[41] Tronto, J. C. (1993). Moral Boundaries: A Political Argument for an Ethic of Care.
[42] Vallor, S. (2016). Technology and the Virtues. Oxford University Press.
[43] World Economic Forum. (2024). AI Value Alignment: Guiding Artificial Intelligence Towards Shared Human Goals. Dignum, V. (2024). AI value alignment: guiding artificial intelligence towards shared human goals. White paper.
[44] Yan, Y., Liu, H., & Chau, T. (2025). A systematic review of AI ethics in education: Challenges, policy gaps, and future directions. Journal of Global Information Management, 33(1), 1–50.
[45] Zhang, Y., Suresh, H., & Netter, J. (2025). Designing stakeholder-based pedagogy for AI ethics education: Insights from a multi-institutional case study. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Vol. 8, No. 3, pp. 2785–2795. https://doi.org/10.1609/aies.v8i3.36759.