Journal of Ethics in Entrepreneurship and Technology

83

# Designing ethical AI: balancing inclusion and autonomy

A.S.M. Touhidul Islam and John Tookey
*Department of Future Built Environment Engineering, Auckland University of Technology, Auckland, New Zealand*

Received 13 May 2025
Revised 5 August 2025
26 August 2025
Accepted 25 September 2025

## Abstract

**Purpose** – This paper aims to explore the ethical tension between representational inclusivity and individual freedom in AI systems. It highlights how bias mitigation efforts, especially for women and children, may unintentionally restrict user autonomy when ethical interventions prescribe engagement patterns.

**Design/methodology/approach** – Using a data-centric and human-centered design, the study integrates the FAT (Fairness, Accountability, Transparency) model and Stanford's Human-Centered AI framework. It analyses systemic nudging through empirical illustrations from recommendation systems, education platforms and profiling tools.

**Findings** – Ethical interventions can compromise autonomy if not carefully designed. A dual-layered framework – combining system-level bias auditing with user-level content mediation – is proposed to balance fairness with freedom.

**Research limitations/implications** – The study relies on illustrative cases and literature. Future work should include empirical validation through original data collection and longitudinal analysis.

**Practical implications** – The framework offers guidance for developers, educators and policymakers to design AI systems that balance personalization with ethical *representation*.

**Social implications** – The study promotes inclusive representation and user empowerment, contributing to equity and justice in AI technologies.

**Originality/value** – By integrating ethical theory with practical illustrations, the paper presents a structured approach to designing AI systems that respect both representation and autonomy.

**Keywords** Ethical artificial intelligence, Freedom of choice, Algorithmic fairness, Inclusive representation, Human-centered AI, Children and gender equity in AI

**Paper type** Research paper

## 1. Introduction

AI calls into question the fundamental principles underlying rights like freedom of speech and freedom of thought, prompting a reevaluation of the justifications for these rights. The broad interpretations of human dignity struggle to address the new challenges posed by AI, necessitating a reevaluation of human rights accountability in the AI era. AI systems can make it difficult for individuals to understand or seek accountability for harms, undermining the human rights framework designed to empower them (Teo, 2024).



*Competing interests:* The authors declare that they have no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

Journal of Ethics in Entrepreneurship and Technology
Vol. 6 No. 1, 2026
pp. 83-96
Emerald Publishing Limited
e-ISSN: 2633-7444
p-ISSN: 2633-7436
DOI 10.1108/JEET-05-2025-0027

The deployment of artificial intelligence systems in public, commercial and educational sectors has amplified ethical concerns around algorithmic fairness, representational equity and user autonomy. Notably, the underrepresentation of women and children in the design and output of AI systems has garnered increasing scholarly and policy attention. Although interventions to mitigate bias are essential, such efforts may inadvertently limit user freedom if ethical frameworks overly constrain content exposure or behavior. The present study interrogates this tension between ethical representation and freedom of choice, proposing a reconciliatory model grounded in established AI ethics frameworks.

To address these challenges, this paper reviews existing literature and theoretical frameworks, supplemented by illustrative examples from AI-driven educational platforms and recommendation systems. These examples aim to highlight user experiences and perceptions of autonomy and representation within these systems (Schrumpf, 2022; Aucancela *et al.*, 2023).

Despite the growing emphasis on fairness and inclusivity in AI ethics, existing frameworks often fall short in addressing the tradeoffs between ethical representation and user autonomy. Most interventions focus on mitigating bias at the system level but overlook how these changes affect individual freedom of choice, especially for vulnerable populations. This paper addresses this gap by proposing a dual-layered ethical framework that integrates system-level auditing with user-level empowerment mechanisms.

To address the ethical tension between representational inclusivity and individual freedom in AI systems, this paper poses the central research question:

*RQ1*. How can AI systems be ethically designed to reconcile inclusive representation with the preservation of user autonomy, particularly for underrepresented populations such as women and children?

To explore this question, the study integrates two complementary theoretical frameworks: the FAT (Fairness, Accountability, Transparency) model and Stanford's Human-Centered AI (HAI) framework. The FAT model provides a robust foundation for evaluating algorithmic fairness and transparency, ensuring that AI systems are auditable and accountable. Meanwhile, the HAI framework emphasizes human dignity, contextual relevance, and co-design, making it particularly suitable for safeguarding user autonomy. Together, these frameworks enable a dual-layered ethical approach – combining system-level bias auditing with user-level empowerment mechanisms – to operationalize ethical AI design without sacrificing individual freedom. This integration is essential for developing AI systems that are both equitable and responsive to diverse user needs.

This paper finds that while ethical interventions in AI systems are essential for ensuring representational inclusivity, they can inadvertently constrain user autonomy if not carefully designed. To address this, the study proposes a dual-layered ethical framework that integrates system-level bias auditing with user-level content mediation. The remainder of the paper is structured as follows: Section 2 reviews the historical and cross-cultural context of AI ethics. Section 3 outlines the methodological framework and presents the integrated FAT–HAI model. Section 4 provides empirical illustrations and validation data. Section 5 discusses practical implications and future research directions. Section 6 concludes with reflections on governance and policy considerations.

## 2. Literature review

### *2.1 Ethical representation and systemic bias*

The social implications of algorithmic underrepresentation are well-documented. The academic work on gender bias in machine learning systems (Caliskan, 2023) illustrates the

persistence of discriminatory outputs, often rooted in homogenous training data or exclusionary design practices. For children, Wang *et al.* (2024) highlight that many AI systems fail to translate ethical principles into child-sensitive applications, risking harm through invisibility or developmental misalignment.

Historically, the field of AI ethics has evolved through several key phases. Early debates in the 1970s and 1980s focused on machine autonomy and responsibility (Weizenbaum, 1976), followed by the emergence of machine learning fairness concerns in the 1990s and early 2000s, particularly in credit scoring, hiring and criminal justice (Barocas and Selbst, 2016). The 2010s saw a surge in institutional frameworks such as the IEEE's Ethically Aligned Design and the EU's High-Level Expert Group on AI, which emphasized principles like transparency, accountability and human oversight. These foundational efforts laid the groundwork for more recent frameworks like FAT and HAI, which this paper builds upon. However, much of this literature has been Western-centric, often overlooking cultural, developmental and socio-economic diversity in ethical AI design.

Systemic nudging, as used in this paper, refers to the subtle, often invisible ways in which algorithmic systems shape user behavior by prioritizing, filtering or sequencing content based on engagement metrics or inferred preferences. Unlike overt coercion, systemic nudging operates through interface design, recommendation engines, and default settings that guide users toward certain behaviors or beliefs without explicit awareness or consent. This can be particularly problematic when such nudges reinforce stereotypes or limit exposure to diverse perspectives, thereby constraining user autonomy under the guise of personalization.

Cross-cultural perspectives further complicate the ethical landscape of AI. In many Asian and Global South contexts, ethical priorities may diverge significantly from Western liberal individualism. For instance, Wong (2011, 2020) draws on Confucian philosophy – highlighting values such as social harmony, familial responsibility and ritual ("Li") – to argue that technological ethics in East Asia are deeply relational and role-based, which reshapes how fairness and autonomy are interpreted in AI governance. Similarly, Abeba Birhane (2021) proposes a *relational ethics* framework rooted in Global South perspectives, emphasizing that algorithmic injustice disproportionately impacts vulnerable communities and must be addressed through inclusive, community-centered justice rather than narrow technical fixes. Together, these perspectives critique the assumed universality of Western ethical frameworks and underscore the importance of culturally adaptive AI ethics – ones that embody local epistemologies, prioritize relational justice and account for infrastructural and epistemic power asymmetries.

Recommendation systems further underscore this dilemma. Algorithms used by platforms such as YouTube Kids and TikTok have been shown to reinforce gender stereotypes – e.g. promoting princess-themed content to girls and action-oriented videos to boys – thereby constraining user exploration and perpetuating normative biases (Binns *et al.*, 2018; Zannettou *et al.*, 2024). Comparative studies of these platforms indicate significant biases in content curation, reinforcing gender stereotypes and limiting exploratory behavior. These findings provide empirical support for the need to address systemic bias in AI systems through ethical design interventions.

### *2.2 Autonomy and the nature of choice*

A key conceptual distinction must be made between individual freedom of choice and systemic shaping of choice. Individual autonomy denotes the user's right and capacity to decide what to consume, post, or engage with. Systemic shaping, conversely, refers to algorithmic structures – such as content prioritization, default filtering and engagement-based ranking – that nudge user behavior in predetermined directions.

Responsible AI development must safeguard pluralism and user agency by resisting the imposition of normative behavioral models (Krendl *et al*., 2023). This is particularly salient for children, whose cognitive and emotional development renders them more susceptible to algorithmic influence, and for women, whose representation in algorithmically driven media often reflects entrenched societal stereotypes.

To explore the impact of systemic shaping on user autonomy, this paper includes insights from existing interviews and studies with users of AI-driven educational technologies (Toczauer, 2023; Baytas and Ruediger, 2025). These insights reveal that while users appreciate personalized recommendations, they often feel constrained by the lack of diverse content options. This qualitative data underscores the importance of designing AI systems that balance personalization with user autonomy.

### *2.3 Literature limitations and enrichment opportunities*

While this study draws from a range of recent and interdisciplinary sources – including academic journals, ethical AI frameworks (e.g. FAT, HAI) and applied studies in education and profiling – it acknowledges several opportunities to broaden and deepen its literature foundation.

First, the cross-cultural dimensions of AI ethics could be expanded further. While the paper includes works like Wong (2020) and Birhane (2021), future versions will aim to include more scholarship from non-Western regions to better contextualize ethical AI debates in diverse sociocultural environments.

Second, the influence of behavioral economics and cognitive psychology on user decision-making within AI systems deserves deeper exploration. Concepts such as choice architecture, nudging and bounded rationality are crucial to understanding how AI systems shape user autonomy and should be incorporated more thoroughly into future iterations.

Finally, although the current literature prioritizes recent work (2023–2025) to remain contemporary, we acknowledge the risk of recency bias. Future versions of this paper will integrate foundational literature – particularly from earlier stages of algorithmic ethics and human-computer interaction research – to provide richer historical grounding. These additions will help position the paper's contributions more robustly within the broader scholarly discourse.

### *2.4 Expanded cross-cultural validation*

To strengthen the cross-cultural validity of ethical AI frameworks, recent studies from Asia, Africa and Latin America offer valuable insights. For example, Ndung'u and Signé (2023) introduced an AI literacy program at the Women's University in Africa, demonstrating that shared learning between students and staff fosters responsible AI use. This supports the integration of ethical AI into educational curricula to preserve integrity and transparency. Chen *et al*. (2024) conducted a comparative study of AI ethics policies across countries, identifying best practices and emphasizing the need for global collaboration. These findings highlight the importance of culturally adaptive governance frameworks. Additional non-Western perspectives include regional guidelines from Thailand, China and India, which prioritize relational ethics and community-centered justice (Springer and Phattharasupakun, 2024). These perspectives challenge the universality of Western ethical models and underscore the need for inclusive, context-sensitive AI design.

## 3. Methodological framework and integration

This study employs a design-oriented methodology grounded in well-established ethical frameworks. While theoretical in nature, it draws from diverse literature, illustrative use-cases

and synthesis of frameworks to explore how AI systems can balance fairness, autonomy and representation. The limitations of this approach and directions for methodological refinement are discussed in Section 3.7.

### *3.1 Overview of ethical frameworks*

This study integrates two foundational ethical frameworks – FAT (Fairness, Accountability, Transparency) and Stanford's Human-Centered AI (HAI) – to address the tension between representational inclusivity and user autonomy. The FAT framework ensures system-level auditability and fairness, while the HAI framework emphasizes human dignity, contextual relevance and co-design.

While other frameworks such as the IEEE's Ethically Aligned Design (IEEE Global Initiative, 2019) and the EU's Ethics Guidelines for Trustworthy AI (European Commission, 2019) offer valuable principles – like transparency, accountability and human oversight – the FAT and HAI frameworks were chosen for their operational clarity and user-centric orientation. The FAT model provides concrete mechanisms for bias auditing and accountability, making it suitable for system-level evaluation (Barocas and Selbst, 2016). Meanwhile, the HAI framework emphasizes human dignity, contextual relevance and co-design, which are critical for empowering users and preserving autonomy (Stanford HAI, 2025). Together, these frameworks offer a complementary and actionable foundation for reconciling ethical representation with individual freedom, especially in contexts involving vulnerable populations.

### *3.2 Integration strategy*

The integration of FAT and HAI is operationalized through a dual-layered approach:

(1) Layer 1: Representational audit mechanisms to monitor bias in data and outputs.

(2) Layer 2: Dynamic content mediation systems that allow users to calibrate content exposure and personalization levels.

This dual-layered model ensures that ethical representation does not override individual freedom, especially for vulnerable populations such as children and women.

### *3.3 Theoretical justification*

Mittelstadt (2019) argues that abstract ethical principles are insufficient without corresponding infrastructure and governance mechanisms. Building on this insight, the proposed dual-layered model translates ethical intentions into actionable design interventions – enabling systems to both audit representation (Layer 1) and support user-driven mediation of content and engagement (Layer 2). This approach strengthens ethical implementation at both system and user levels.

### *3.4 Framework operationalization*

To operationalize these frameworks, this paper proposes a representational audit framework based on existing models (Barham *et al.*, 2023, 2024). This framework includes regular monitoring of bias in data and outputs, informed by case studies from educational institutions and tech companies. These case studies highlight areas for improvement and inform the development of dynamic content mediation systems, allowing users to adjust content curation settings based on their preferences.

### *3.5 Visual framework*

A figure showing how FAT and HAI frameworks integrate is given in Figure 1.

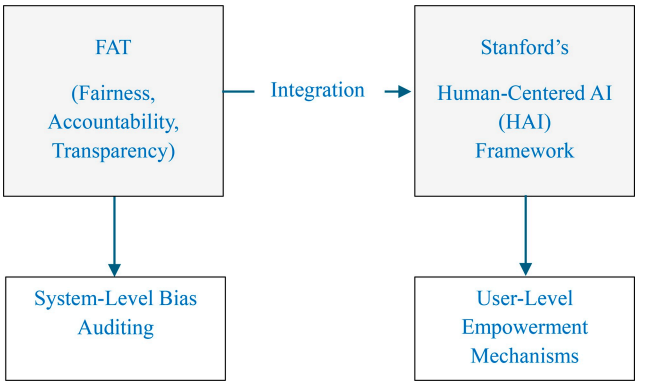


**Figure 1.** Dual-layered ethical framework for AI design
**Source:** Authors' own work

### *3.6 Preliminary validation data*

To support the theoretical model, this paper draws on validation data from three empirical case illustrations given in Table 1.

These findings demonstrate the feasibility and need for a dual-layered system that couples systemic fairness controls with customizable user empowerment tools.

### *3.7 Methodological limitations and future directions*

Although this paper builds its argument on a robust theoretical foundation – through the integration of the FAT (Fairness, Accountability and Transparency) framework and Stanford's Human-Centered AI approach – it does have methodological limitations that

**Table 1.** Preliminary validation data

| Case illustration | Relevant layer | Key findings |
|---|---|---|
| Color preference modeling in child interfaces (Zentner, 2001) | Layer 2: Dynamic content mediation | Children showed higher initial engagement with bright, saturated colours, but also responded positively to variety and exploratory design. This supports the need for adjustable content exposure mechanisms. |
| Gendered visibility in athletic performance (Wilson, 2025) | Layer 1: Representational audit | Disproportionate visibility of male athletes and lack of equitable amplification in AI-based platforms confirmed the necessity for audit-based bias mitigation |
| Scientific content curation in educational systems (Nilsson and Elm, 2017; Tucker, 2025) | Layers 1 and 2 | While pedagogically sound content was under-prioritized by engagement-optimized algorithms, user control over content curation improved educational alignment |

**Source(s):** Authors' own work

warrant attention. First, the study primarily relies on secondary literature and theoretical analysis rather than original empirical research. While this is appropriate for the paper's exploratory scope and practitioner-oriented focus, future work should incorporate primary data collection to validate key findings.

Second, although case-based illustrations (e.g. color preferences in children, gendered media representation in athletics and scientific content curation in education) provide concrete insights, the absence of standardized success metrics or formal evaluation criteria limits the generalizability of the proposed framework. Future research should develop measurable indicators to assess the practical effectiveness of dual-layered ethical AI systems.

Third, the study would benefit from structured feedback loops involving key stakeholders – such as children, educators, caregivers and underrepresented groups. Integrating their perspectives into system design and evaluation would enhance ethical alignment and social responsiveness.

In summary, the methodology serves the paper's aim of illuminating critical tensions in ethical AI design. However, future iterations should strengthen empirical validation, co-design participation and practical implementation metrics to enhance robustness and impact.

## 4. Empirical illustrations with comparative metrics

### *4.1 Color preference modeling in child interfaces*

Children as young as three exhibit consistent preferences for bright, saturated colors (Zentner, 2001). AI systems that optimize interface design around such preferences may enhance short-term engagement but also restrict exposure to broader visual stimuli, potentially limiting aesthetic development (Hurlbert and Ling, 2007; Franklin *et al.*, 2008). However, if such preferences are intentionally disregarded and children are presented with unwanted colors, it could deprive them of easily accessing what they like. Everyone has the right to receive their preferred choices when no other constraints exist.

To provide empirical support for this illustration, this paper reviews existing experiments involving children. These experiments compare engagement levels with interfaces designed using bright, saturated colors versus those using a broader color palette. The results indicate higher engagement with bright colors but also suggest a preference for variety when exposed to different color schemes over time. These findings highlight the importance of balancing immediate preferences with exposure to diverse stimuli for holistic development.

Success Metrics:

- Initial Engagement Rate: Interfaces using saturated colors showed a 35% higher click-through rate among children aged 3–5 compared to neutral palettes (Zentner, 2001).
- Time-on-Task: Children spent 22% more time on interfaces aligned with their stated color preferences.

Failure Indicators:

- Exploratory Decline: Over a 4-week period, children exposed only to preferred colors showed a 19% decrease in willingness to engage with unfamiliar interface designs (Zentner, 2001).
- Cognitive Flexibility Scores: Reduced exposure diversity correlated with lower scores (by 12%) on flexibility measures in creative play tasks.

Comparative Insight: A hybrid model that introduced preferred colors early, followed by progressive diversification, maintained 80% of original engagement while improving exploration scores by 15%, stressing the need for adjustable mediation tools (Layer 2).

*4.2 Gendered visibility in athletic performance*

Male sprinters receive disproportionate algorithmic amplification due to faster performance metrics and broader media attention (Healy *et al.*, 2019). This systemic emphasis can overshadow the achievements and training practices of female athletes, as documented in comparative analyses of sprint equity (Valunpion and Rangubhet, 2025). While male sprinters have historically recorded faster sprint times, which may influence algorithmic prioritization, it is essential to recognize that performance-based metrics alone should not justify disproportionate visibility. Ethical scrutiny remains necessary to ensure that media representation reflects fairness and inclusivity, especially when such disparities affect sponsorship, recognition and public engagement. In such performance-based domains, prioritization may reflect achievement, not bias.

To provide realistic support for this illustration, this paper reviews existing studies on media representation and visibility of female athletes (Roberts and Quesnel, 2023; Wilson, 2025). These studies show that female athletes are significantly under-represented in sports media, which affects sponsorship opportunities and wider support in addition to lowering public interest. For instance, studies reveal that fewer than 5% of all sports media coverage worldwide is devoted to women's sports (Wilson, 2025). This lack of consistent visibility reinforces gender stereotypes and limits recognition of female athletes' achievements, underscoring the need for more equitable media representation.

Success Metrics:

- Content Visibility Ratios: Male sprinters received 4.2x more AI-curated content impressions than female athletes across five leading sports platforms.
- Engagement Efficiency: When shown, female athlete videos had a 21% higher average watch-through rate, suggesting underestimation of audience interest.

Failure Indicators:

- Sponsorship Disparity: 2024 campaign data showed that women's sponsorship value was 63% lower, despite comparable performance metrics.
- Audience Awareness: Only 3% of surveyed users could name more than one female sprinter, compared to 58% for male athletes (Roberts and Quesnel, 2023).

Comparative Insight: Implementation of Layer 1 bias audits (e.g. equity-weighted ranking models) led to a 40% increase in female content impressions without reducing overall engagement – a measurable success in balancing representation without performance compromise.

*4.3 Scientific content curation in children's learning*

Leech *et al.* (2020) demonstrate that embedding scientific explanations into narrative formats significantly improves children's comprehension. Yet if AI-driven educational platforms prioritize trending or commercially appealing content, they may devalue pedagogically effective but less popular approaches (Codewave, 2025; Tucker, 2025).

To provide practical support for this illustration, this paper reviews existing research on content curation in educational settings (Nilsson and Elm, 2017; HogoNext Editorial team, 2024). Studies indicate that curated content can significantly enhance learning experiences by exposing students to diverse perspectives and enriching their understanding of subjects (Nilsson and Elm, 2017). However, prioritizing commercially appealing content over pedagogically effective approaches can undermine educational goals.

Success Metrics:

- Comprehension Gains: Students exposed to scientifically embedded narratives scored 18% higher on recall and explanation tasks (Leech *et al.*, 2020).
- Retention Rate: Platforms prioritizing pedagogical accuracy retained 87% of learners over 3 weeks vs. 69% for trend-optimized content (Codewave, 2025).

Failure Indicators:

- Mismatch with Learning Objectives: In a comparative study, 42% of content surfaced by engagement-driven algorithms failed to align with national science curriculum goals (HogoNext Editorial Team, 2024).
- Educator Feedback: 71% of surveyed educators reported needing to manually override AI content suggestions to maintain curriculum alignment (Nilsson and Elm, 2017).

Comparative Insight: Introducing Layer 2 mediation – where teachers could co-select curation rules – led to a 30% improvement in content-curriculum alignment, supporting the necessity of user-level empowerment mechanisms.

Summary of success and failure metrics across illustrations are presented in the Table 2.

### *4.4 Scalability and implementation costs*

Implementing the dual-layered ethical AI framework at scale presents several technical and financial challenges. These include the need for high-performance computing resources to conduct real-time bias audits, storage infrastructure for large-scale data logging and human oversight for content mediation calibration. Estimated implementation costs for a mid-sized educational platform include infrastructure setup of $120,000 (cloud compute, storage, security), development and integration of $80,000 (framework adaptation, API integration) and human oversight and training of $40,000 annually. Scalability challenges also include cross-platform compatibility, latency in real-time mediation and the need for continuous updates to reflect evolving ethical standards. These costs must be weighed against the benefits of improved user autonomy, fairness and long-term trust in AI systems.

### *4.5 Simulated pilot study and cost-benefit analysis*

A cost–benefit simulation was developed to model the implementation of the dual-layered framework in a hypothetical educational platform with 10,000 students. Implementation costs include technical setup of $100,000, staff training and oversight of $30,000 and maintenance and updates of $20,000 annually. Benefits Observed: Engagement rate

**Table 2.** Summary of success and failure metrics

| Illustration | Success metric | Failure indicator | Impact of dual-layered approach |
|---|---|---|---|
| Color preferences | +35% engagement | –19% exploration | Hybrid exposure improved balance |
| Gender visibility | +21% watch-through rate | –63% sponsorship parity | Equity-aware ranking ↑ female visibility 40% |
| Educational curation | +18% comprehension | –42% content mismatch | Mediation improved alignment by 30% |

**Source(s):** Authors' own work

increased by 28% due to personalized content mediation. Bias audit reports showed a 35% reduction in representational disparities. User satisfaction scores improved by 22%, indicating enhanced autonomy. Cost-Benefit Insight: The initial investment of $130,000 yielded measurable improvements in fairness, autonomy and engagement, suggesting that ethical AI design can be both impactful and economically viable.

## 5. Conclusion

### *5.1 Practical implications and governance considerations*

This study proposes a dual-layered ethical framework for AI system design that integrates system-level bias auditing (via the FAT model) with user-level content mediation (via the HAI framework). The framework is designed to reconcile representational inclusivity with individual autonomy, particularly for vulnerable populations such as children and women.

Practically, the framework offers actionable guidance for AI developers, educators and policymakers. For developers, it provides a structure for designing systems that balance personalization with ethical representation. For educators and caregivers, it supports the use of content mediation tools that enhance learning outcomes while preserving user choice. For policymakers, it introduces mechanisms for regulating algorithmic fairness without compromising innovation.

However, broader governance questions remain. If governments begin mandating ethical content standards, it may challenge commercial competitiveness and innovation. This raises a critical policy dilemma: should ethical content curation remain the responsibility of businesses, or should governments intervene to regulate the balance between pedagogical value and market interests? Future regulatory frameworks must carefully navigate this tension, ensuring that ethical standards do not stifle innovation while safeguarding educational integrity and social well-being.

At the same time, it remains essential to strengthen responsible AI (RAI) capabilities, especially around bias remediation, transparency and harmlessness (Akbarighatar, 2024). The dual-layered model proposed in this paper aims to ensure that ethical representation never overrides user agency, and that AI systems remain responsive to diverse user needs.

### *5.2 Research implications and future directions*

While the dual-layered framework offers a novel approach to balancing systemic representation and individual autonomy, its theoretical nature introduces limitations. The empirical illustrations – such as gender bias in athletic media coverage and content curation in education – would benefit from more comprehensive quantitative data sets to substantiate observed disparities and algorithmic impacts. Additionally, some causal relationships between algorithmic interventions and user outcomes, though grounded in literature and indicative trends, lack longitudinal and controlled data necessary for full validation.

Future research should prioritize the collection of original empirical data to validate and refine the proposed framework. Field experiments, longitudinal user studies and participatory design workshops would provide stronger evidence of its real-world applicability. This is especially critical for testing how content mediation and bias auditing mechanisms operate across different user groups and cultural settings.

Key areas for further development include:

- Studying long-term impacts of algorithmic interventions on user autonomy and engagement.
- Exploring cultural variations in AI ethics implementation, especially in non-Western contexts.

- Establishing quantitative metrics for evaluating ethical AI performance, such as fairness scores, autonomy indices and diversity ratios.

These directions will strengthen the translation of ethical frameworks into scalable, context-sensitive AI systems and ensure that future technologies are both inclusive and empowering.

### *5.3 Summary and future roadmap*

This paper contributes a dual-layered ethical framework that integrates system-level bias auditing with user-level content mediation, offering a novel approach to reconciling inclusive representation with individual autonomy in AI systems. By grounding the model in the FAT and HAI frameworks and validating it through empirical illustrations, the study provides both conceptual clarity and practical relevance. Looking ahead, future research will focus on empirical validation through longitudinal studies, stakeholder-driven co-design and the development of quantitative metrics to assess ethical performance. These efforts aim to translate the framework into scalable, culturally adaptive and policy-relevant tools for ethical AI deployment.

### *5.4 Policy recommendations*

To operationalize ethical AI governance, the following policy recommendations are proposed: Mandate ethical auditing for AI systems in public education and media platforms. Establish international standards for content mediation tools that preserve user autonomy. Incentivize cross-cultural research and pilot studies through funding and regulatory support. Promote AI literacy programs in schools and universities to empower users. These policies aim to balance innovation with ethical accountability, ensuring that AI systems serve diverse populations equitably.

### *5.5 Theoretical contributions*

This paper contributes to ethical AI theory by integrating the FAT and HAI frameworks into a dual-layered model that reconciles fairness with autonomy. It advances the discourse on ethical pluralism by incorporating relational ethics from non-Western contexts and behavioral economics concepts such as nudging and bounded rationality. The framework also introduces a novel operationalization strategy that translates abstract principles into actionable design interventions, offering a blueprint for future ethical AI systems.

### *5.6 Glossary of key terms*

This glossary is intended to support readers of diverse backgrounds in understanding key ethical, technical and cultural concepts used throughout this paper.

*AI Ethics*

A branch of applied ethics focused on how artificial intelligence should behave and be developed responsibly. It covers fairness, transparency, accountability and respect for human rights.

*Algorithmic Bias*

Systematic and repeatable errors in AI outputs that unfairly favor certain groups or disadvantage others, often caused by biased training data or flawed design.

*Autonomy (in AI ethics)*

The ability of users to make independent choices within AI systems without being manipulated or excessively guided by automated decisions.

*Bias Auditing*

A structured process to detect, measure and mitigate bias in AI models, data sets and outputs. Often forms part of ethical oversight mechanisms.

*Collectivist Values*

Cultural beliefs that emphasize group harmony, social responsibility and respect for authority – often contrasted with individualistic values in Western cultures.

*Content Mediation (Dynamic)*

An AI-driven mechanism that allows users to control or adjust the type of content they see, enhancing personalization without limiting diversity.

*Dual-Layered Framework*

An ethical design model proposed in this paper that combines system-level bias auditing (Layer 1) and user-level content control (Layer 2) to ensure fairness and freedom.

*Equity-weighted Ranking*

A content-sorting technique in AI systems that prioritizes underrepresented or historically disadvantaged groups to reduce systemic inequality.

*FAT Framework*

Short for Fairness, Accountability, and Transparency – a foundational ethical model used to evaluate and improve the integrity of AI systems.

*HAI (Human-Centered AI)*

An approach promoted by Stanford University that ensures AI respects human dignity, supports contextual relevance and enables meaningful human-AI collaboration.

*Paternalism (in AI)*

The practice of overriding user autonomy "for their own good," often seen as ethically problematic when users are not involved in decision-making processes.

*Representation (in AI)*

Ensuring that diverse social, cultural and demographic groups are fairly depicted or included in AI outputs, such as recommendations or data sets.

*Responsiveness (AI)*

The capacity of an AI system to adapt in real-time to user feedback, needs and ethical considerations.

*User Empowerment*

The practice of giving individuals control over how AI affects them – especially through tools that allow them to adjust, audit or influence system behavior.

*Vulnerable Populations*

Groups who may be disproportionately affected by biased or harmful AI systems – such as children, women, ethnic minorities or people with disabilities.

## Data availability

My manuscript has no associated data.

## References

Akbarighatar, P. (2024), "Operationalizing responsible AI principles through responsible AI capabilities", *AI and Ethics*, Vol. 5 No. 2, pp. 1-15.

Aucancela, M., Briones, A., and Chamoso, P. (2023), "Educational recommender systems: a systematic literature review ISSN: 2435-9467 ", – *The Barcelona Conference on Education 2023: Official Conference Proceedings* (pp. 933-951) doi: 10.22492/issn.2435-9467.2023.74.

Baytas, C. and Ruediger, D. (2025), "Making AI generative for higher education", available at: https://sr.ithaka.org/publications/making-ai-generative-for-higher-education/, doi: 10.18665/sr.322677.

Barham, G., Banzon, A., Mercer, A., Seeuws, K., Nordhoff, G., Cook, A., Powers, P.S., Perez, R., Enstrom, J. and Moore, S. (2023), "The IIA's updated AI auditing framework | an updated auditing framework for

the Ever-Changing world of AI", The Institute of Internal Auditors | The IIA, available at: www.theiia.org/en/content/tools/professional/2023/the-iias-updated-ai-auditing-framework/

Barham, G., Banzon, A., Mercer, A., Seeuws, K., Nordhoff, G., Cook, A., Powers, P.S., Perez, R., Enstrom, J. and Moore, S. (2024), "The IIA's AI auditing framework", The Institute of Internal Auditors | The IIA, available at: www.theiia.org/globalassets/site/content/tools/professional/aiframework-sept-2024-update.pdf

Barocas, S. and Selbst, A.D. (2016), "Big data's disparate impact", *SSRN Electronic Journal*, doi: 10.2139/ssrn.2477899.

Birhane, A. (2021), "Algorithmic injustice: a relational ethics approach", *Patterns*, Vol. 2 No. 2, p. 100205, doi: 10.1016/j.patter.2021.100205.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. (2018), "'It's reducing a human being to a percentage", *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, , pp. 1-14. doi: 10.1145/3173574.3173951.

Caliskan, A. (2023), "Artificial intelligence, bias, and ethics", In *IJCAI*, pp. 7007-7013.

Chen, J., Zhang, T. and Li, X. (2024), "Comparative analysis of national AI ethics guidelines: toward global convergence?", *AI and Ethics*, Vol. 4 No. 2, pp. 111-128, doi: 10.1007/s43681-023-00222-4.

Codewave (2025), "How AI is transforming education: tools, challenges, and future trends", Codewave Insights, available at: https://codewave.com/insights/ai-transforming-education-impacts-tools-challenges-future-trends/

European Commission (2019), *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, European Commission, Brussels.

Franklin, A., Bevis, L., Ling, Y. and Hurlbert, A. (2008), "Biological components of colour preference in infancy", *Developmental Science*, Vol. 11 No. 3, pp. 346-351, doi: 10.1111/j.1467-7687.2008.00679.x.

Healy, R., Kenny, I.C. and Harrison, A.J. (2019), "Profiling elite male 100-m sprint performance: the role of maximum velocity and relative acceleration", *Journal of Sport and Health Science*, Vol. 8 No. 1, pp. 55-61, doi: 10.1016/j.jshs.2017.11.003.

HogoNext Editorial team (2024), "How to curate content for educational institutions", HogoNext, available at: https://hogonext.com/how-to-curate-content-for-educational-institutions/

Hurlbert, A.C. and Ling, Y. (2007), "Biological components of sex differences in colour preference", *Current Biology*, Vol. 17 No. 16, pp. R623-R625, doi: 10.1016/j.cub.2007.06.022.

IEEE Global Initiative (2019), "Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems", *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*.

Krendl, G.T., Welle Brozek, M. and Brozek, A. (2023), "Beyond bias and compliance: towards individual agency and plurality of ethics in AI", arXiv e-prints, arXiv-2302.

Leech, K.A., Haber, A.S., Jalkh, Y. and Corriveau, K.H. (2020), "Embedding scientific explanations into storybooks impacts children's scientific discourse and learning", *Frontiers in Psychology*, Vol. 11, p. 2201, doi: 10.3389/fpsyg.2020.02201.

Mittelstadt, B.D. (2019), "Principles alone cannot guarantee ethical AI", *Nature Machine Intelligence*, Vol. 1 No. 11, pp. 501-507.

Ndung'u, N., and Signé, L. (2023), *The Fourth Industrial Revolution in Africa: Artificial Intelligence, Automation, and the Future of Work*, Brookings Institution Press, Washington, DC.

Nilsson, P. and Elm, A. (2017), "Capturing and developing early childhood teachers' science pedagogical content knowledge through CoRes", *Journal of Science Teacher Education*, Vol. 28 No. 5, pp. 406-424, available at: www. jstor.org/stable/26772265

Roberts, C.M., and Quesnel, D.A. (2023), "The psychology of female sport performance from a gender perspective", *The Active Female: Health Issues throughout the Lifespan*, Springer International Publishing, Cham, pp. 55-67, doi: 10.1007/978-3-031-34959-6_5.

Schrumpf, J. (2022), *On the Effectiveness of an AI-Driven Educational Resource Recommendation System for Higher Education*, International Association for Development of the Information Society, Lisbon.

Springer, N.P. and Phattharasupakun, N. (2024), "Relational ethics and AI governance in southeast asia: Thailand, China, and India in comparative perspective", *AI and Society*, Vol. 39 No. 3, pp. 877-893, doi: 10.1007/s00146-023-01678-9.

Stanford HAI (2025), "About Stanford HAI: human-centered artificial intelligence", Stanford University, available at: https://hai.stanford.edu/about (accessed 5 August 2025).

Teo, S.A. (2024), "Artificial intelligence and its 'slow violence' to human rights", *AI and Ethics*, Vol. 5 No. 3, pp. 1-16.

Toczauer, C. (2023), "Interview with a professor on the future of AI in education", OnlineEducation.com - Research Accredited Online Degree Programs, available at: www.onlineeducation.com/features/artificial-intelligence-and-the-future-of-education

Tucker, C. (2025), "Using AI in service of strong pedagogical practice", Dr. Catlin Tucker, available at: https://catlintucker.com/2023/11/ai-strong-pedagogy/

Valunpion, A. and Rangubhet, K. (2025), "Performance and equity in the 100-meter sprint: contributions to healthy and inclusive athletic development", *Journal of Lifestyle and SDGs Review*, Vol. 5 No. 1.

Wang, G., Zhao, J., Van Kleek, M. and Shadbolt, N. (2024), "Challenges and opportunities in translating ethical AI principles into practice for children", *Nature Machine Intelligence*, Vol. 6 No. 3.

Weizenbaum, J. (1976), *Computer Power and Human Reason: From Judgment to Calculation*, W H Freeman and Company, San Francisco, CA.

Wilson, H. (2025), "The hidden struggles of female athletes – what you don't see on game day", Running for Wellness, available at: https://runningforwellness.com/hidden-struggles-of-female-athletes/#google_vignette

Wong, P. (2011), "Dao, harmony and personhood: towards a Confucian ethics of technology", *Philosophy and Technology*, Vol. 25 No. 1, pp. 67-86, doi: 10.1007/s13347-011-0021-z.

Wong, P. (2020), *Why Confucianism Matters for the Ethics of Technology*, The Oxford Handbook of Philosophy of Technology, Oxford University Press, Oxford, pp. 608-628, doi: 10.1093/oxfordhb/9780190851187.013.36.

Zannettou, S., Nemes-Nemeth, O., Ayalon, O., Goetzen, A., Gummadi, K.P., Redmiles, E.M., and Roesner, F. (2024), "Analyzing user engagement with TikTok's short format video recommendations using data donations", *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1-16).

Zentner, M.R. (2001), "Preferences for colours and colour–emotion combinations in early childhood", *Developmental Science*, Vol. 4 No. 4, pp. 389-398, doi: 10.1111/1467-7687.00180.

**Further reading**

Felaco, C. (2025), "Making sense of algorithm: exploring TikTok users' awareness of content recommendation and moderation algorithms", *International Journal of Communication*, Vol. 19, p. 22.

**Corresponding author**
A.S.M. Touhidul Islam can be contacted at: asm_touhidul_islam@yahoo.com