6-2024

Navigating the ethical terrain of AI in higher education: Strategies for mitigating bias and promoting fairness

Emily Barnes

James Hutson

Forum for Education Studies 2024, 2(2), 1229.
https://doi.org/10.59400/fes.v2i2.1229

Article

Navigating the ethical terrain of AI in higher education: Strategies for mitigating bias and promoting fairness

Emily Barnes*, **James Hutson**

Lindenwood University, St Charles, MO 63301, USA
*** Corresponding author:** Emily Barnes, ejbarnes035@gmail.com

CITATION

Barnes E, Hutson J. Navigating the ethical terrain of AI in higher education: Strategies for mitigating bias and promoting fairness. Forum for Education Studies. 2024; 2(2): 1229.
https://doi.org/10.59400/fes.v2i2.1229

ARTICLE INFO

Received: 21 March 2024
Accepted: 28 April 2024
Available online: 21 June 2024

Abstract: Artificial intelligence (AI) and machine learning (ML) are transforming higher education by enhancing personalized learning and academic support, yet they pose significant ethical challenges, particularly in terms of inherent biases. This review critically examines the integration of AI in higher education, underscoring the dual aspects of its potential to innovate educational paradigms and the essential need to address ethical implications to avoid perpetuating existing inequalities. The researchers employed a methodological approach that analyzed case studies and literature as primary data collection methods, focusing on strategies to mitigate biases through technical solutions, diverse datasets, and strict adherence to ethical guidelines. Their findings indicate that establishing an ethical AI environment in higher education is imperative and involves comprehensive efforts across policy regulation, governance, and education. The study emphasizes the significance of interdisciplinary collaboration in addressing the complexities of AI bias, highlighting how policy, regulation, governance, and education play pivotal roles in creating an ethical AI framework. Ultimately, the paper advocates for continuous vigilance and proactive strategies to ensure that AI contributes positively to educational settings, stressing the need for robust frameworks that integrate ethical considerations throughout the lifecycle of AI systems to ensure their responsible and equitable use.

Keywords: artificial intelligence; ethical challenges; bias mitigation; higher education; algorithmic fairness

1. Introduction

Artificial intelligence (AI) and machine learning (ML) are revolutionizing the realm of higher education, bringing an array of potential and diverse applications that cater to the unique challenges of online learning [1–3]. These rapidly evolving technologies have greatly enriched the teaching and learning experience, offering invaluable support to students and faculty alike and streamlining educational processes [4,5]. Ranging from personalized learning modules and intelligent tutoring systems to automated grading techniques [6–9], these AI-driven tools play a pivotal role in enhancing student success, notably through early recognition systems. Due to this rapid adoption, there is increased need to establish best practices, as Miao et al. [10] have emphasized for teaching and learning with AI and ML.

The successful use of AI in higher education is demonstrated through a variety of innovative applications. Yang et al. [11] created a practical AI-based teaching model that focuses on personalized and cooperative learning, adapting educational experiences to meet the distinct needs of each student. Ali and Abdel-Haq [12] pinpointed important AI uses in academic support, including profiling, prediction, assessment, and adaptive systems, all essential for enhancing student learning experiences [13] and improving the success rates of international students. These recent observations are built on past research, such as that by Popenici and Kerr [14], who explored the rise of AI in higher education, studying its educational impact and the adoption challenges faced by institutions. Their findings stressed the importance of additional research into role of the emergent technology in teaching, learning, and administration to foresee its future influence on higher education. Zawacki-Richter et al. [15] reviewed AI applications in higher education systematically, identifying its application areas while also calling for deeper reflections on its challenges and risks, alongside the limited connection to pedagogical theories. Additionally, Maphosa and Maphosa [16] performed a bibliometric analysis on the impact in higher education, showing worldwide interest but also indicating that scientific literature on AI in this domain is still emerging. These studies collectively emphasize the capacity of the technology to transform higher education and advocate for more comprehensive research, a greater understanding of ethical considerations, and the development of effective teaching strategies.

While integrating AI into higher education offers numerous benefits, it also presents considerable ethical challenges that require careful consideration and action. The rationale for this study stems from the urgent need to address these inherent biases within AI applications in higher education. By exploring how biases manifest in AI-driven educational tools, this study aims to develop strategies to mitigate these biases and promote a fairer learning environment. The research question guiding this study is: How can higher education institutions effectively identify and mitigate inherent biases within AI systems to ensure equitable educational outcomes? This paper aims to present a detailed exploration of the origins, manifestations, and mitigation strategies for biases within AI systems, particularly in the context of higher education. It underscores the critical need for diverse and representative datasets for an effective understanding and addressing of these biases. An interdisciplinary approach is advocated, bringing together insights from technology, ethics, education, and social sciences. This approach is aimed at developing comprehensive strategies to combat biases in AI. The expected significance of this research is in its potential to inform educators, technologists, and policymakers about ethical AI practices, contributing positively and equitably to the future of higher education. The findings of this review are intended to mitigate existing biases in AI and establish a foundation for proactive, inclusive, and ethical development and application of AI in education. This paper serves as a synthesis of existing knowledge and perspectives, aiming to advance the conversation and guide future research and practice in this critical area.

2. Literature review

2.1. The impact of AI on higher education

The integration of AI and ML into higher education is deeply impacted by various biases that can arise both intentionally and unintentionally (**Table 1**). Research has illuminated the repetitive nature of bias in ML algorithms and the long-term implications these biases have on algorithmic performance and the need for inclusivity [17]. For example, gender bias in AI and ML has been a particular focus, with scholars emphasizing the need to integrate diversity and gender theory into these fields to mitigate such biases [18–21]. Further comprehensive surveys by Ntoutsi et al. [17] and Mehrabi et al. [22] have all provided insights into the technical challenges and solutions for bias in AI systems, delving into fairness definitions and mitigation techniques. In particular, Zhou et al. [23] have most recently expanded this discussion by providing a comprehensive view of bias, fairness, and accountability in AI, focusing on supervised ML algorithms.

Table 1. Potential negative impacts of bias in higher education.

Potential negative impact	Description
Perpetuation of Inequality	AI bias can reinforce existing social and educational inequalities, disadvantaging certain groups based on race, gender, socioeconomic status, or other factors [24].
Misguided decision-making	Biased AI may lead to erroneous decisions in admissions, grading, and student support, affecting students' educational and career trajectories [25].
Erosion of trust	The use of biased AI systems can lead to a loss of trust among students, faculty, and stakeholders if they feel the systems are unfair or opaque [26].
Legal and ethical repercussions	Institutions may face legal challenges and ethical criticisms if biased AI systems result in discriminatory practices or violate privacy and fairness norms [27].
Reduction in diversity	If biases in AI are not addressed, they may lead to homogenized student populations and learning environments, reducing the richness and benefits of diverse educational experiences [28].
Hindrance to personalized learning	Biased AI systems may fail to effectively personalize learning experiences, potentially overlooking or misinterpreting individual student needs, preferences, and capabilities due to underlying biases in data or algorithms [29].
Damage to institutional reputation	The use of biased AI can tarnish the reputation of educational institutions, affecting their ability to attract diverse talent, secure funding, and maintain accreditation [30].

2.2. Technical challenges and solutions for bias in AI systems

Mashhadi et al. [31] conducted a case study on integrating fairness visualization tools into ML education. They emphasized the necessity of enhanced education about AI fairness and bias within industry and academia. The study utilized publicly accessible visualization tools to help students explore concepts of algorithmic fairness, shedding light on the advantages, challenges, and potential of such tools in ML education. In a similar vein, Islam et al. [32] integrated fairness and bias themes into an undergraduate computer science curriculum, underlining the critical role of educating future technologists about the risks associated with fairness in AI decision-making systems. Additionally, Borenstein and Howard [33] explored how bias is embedded in contemporary AI and robotic systems. They underscored the importance of implementing strategies to prevent and mitigate bias in these technological domains.

The pervasive issue of bias in AI underscores the critical need for ongoing research, education, and policy development to ensure that AI systems used in higher education are fair, accountable, and transparent. This initial review underscores the multifaceted nature of biases in AI, ranging from gender biases to algorithmic biases, and highlights the need for continued interdisciplinary efforts to address these challenges in the context of higher education. The proactive measures outlined in these studies are instrumental in guiding the development of more equitable AI systems, ensuring that the integration of AI into higher education contributes positively and fairly to the learning experience.

Furthermore, recent case studies from various institutions illustrate the ongoing challenges and initiatives to address and navigate the ethical challenges posed by AI in higher education. At the University of California and Carnegie Mellon University, proactive measures were taken to integrate AI in a fair and ethical manner. Before implementing an algorithm for admissions decisions, the University of California conducted a comprehensive audit to ensure the elimination of potential biases [17]. The development of an AI tutor at Carnegie Mellon underwent a thorough review to ensure that the system did not perpetuate biases related to gender, ethnicity, or learning styles, demonstrating a commitment to creating inclusive educational technologies [34].

2.3. Education and bias mitigation

While these examples demonstrate positive outcomes, others are cautionary in nature. For instance, St. George's University in the United Kingdom faced challenges with its AI system used for screening medical school applicants. After finding that the system was biased against women and certain ethnicities, modifications were made to promote a fair admissions process [15]. Other institutions also employed predictive analytics for student support, necessitating a review of the system to ensure fair treatment of all students, regardless of their demographic background [35–37]. The University of North Carolina used predictive analytics to identify at-risk students, ensuring that the system did not unfairly target certain demographics [38]. Similarly, the Animal Watch AI tutoring system was developed with an overlay student model to adapt problem selection and hints to different learning needs and sensitivities, with observed distinct responses between different genders [39]. These case studies collectively illustrate the diverse manifestations of AI bias in higher education and underscore the critical need for continuous vigilance, proactive measures, and ongoing research to ensure that AI technologies contribute positively to the educational landscape and support an inclusive, equitable environment for all students.

2.4. Societal and ethical implications

The ethical and societal implications of bias in AI within higher education are significant and far-reaching, impacting various aspects beyond just the systems themselves. As AI becomes increasingly integral to diverse educational areas, including mentoring, international student support, and admissions, the importance of ethical principles like fairness, accountability, and transparency is heightened. Ntoutsi et al. [17] offer a comprehensive multidisciplinary perspective on bias in AI systems. Their work emphasizes the technical challenges and potential solutions, and advocates for new research directions rooted in legal frameworks. This underscores the essential need for AI systems to incorporate core values such as transparency, justice, fairness, and privacy throughout their design and deployment processes.

Furthermore, Chu et al. [40] discuss the implications of digital ageism in AI, emphasizing the need to critically examine the presence of age-related bias in AI systems and the broader ethical and legal implications of such biases. The call for a reevaluation of how AI systems are developed and deployed, ensuring they do not perpetuate any form of discrimination, including ageism. The practical ethics of bias reduction in AI systems, as discussed by Tomalin et al. [41], critically evaluates the effectiveness of different strategies for debiasing, particularly in the context of machine translation systems. Such discussions underscore the importance of not only implementing debiasing strategies but also ensuring that these strategies do not compromise the overall performance of the AI systems. Additionally, Li and Xing [42] reveal factors influencing perceived fairness on the part of students towards AI systems in education, which further emphasizes the need to understand and address the perceptions and attitudes of end-users, particularly students, when implementing AI systems in educational settings.

Importantly, bias in AI, particularly in higher education settings, can have profound societal impacts beyond those confined to educational institutions. The ethical and societal implications of bias in AI are substantial, affecting everyone, everywhere, and at any time, raising concerns about potential human rights issues [17]. These biases can perpetuate existing societal inequities, privilege, and power, thus affecting sectors including healthcare, education, employment, and more [40]. Such widespread applications of AI have led to discourse on how these systems are perpetuating racism, sexism, classism, and other forms of discrimination.

Borenstein and Howard [33] delve into how bias has become embedded in current AI and robotic systems, offering specific examples to illustrate this phenomenon. Their analysis highlights how patterns within datasets can carry implicit biases, inadvertently solidifying these biases as universally accepted truths. The authors also contemplate the future design of these systems, proposing strategies to prevent and mitigate the infiltration of bias into robotic technology. Furthermore, they point out that decisions made by these data-driven systems often reflect prejudices based on demographic characteristics such as race and sex. This underscores the urgency of integrating ethical and legal principles into the design, training, and deployment of these systems to promote societal well-being, as emphasized by Ntoutsi et al. [17].

Leavy [20] further discuss the potential of AI to intensify societal biases, potentially reversing progress made in equal rights and civil liberties. They explore ongoing efforts to achieve data justice, fairness, and bias mitigation across various AI system domains. Their work examines how the inherent biases in AI training data could be transformed to serve the social good, highlighting the complex interplay between different dynamics in this process.

The critical discussion around bias in AI encompasses its extensive impact on sectors like employment, education, and financial services. Fu et al. [43] provide an in-depth examination of algorithmic bias, exploring its definition, identification, mitigation, and broader implications. They highlight a significant shift in the perception of machine learning (ML) algorithms, which were once considered neutral but are now increasingly recognized as biased, contributing to structural inequalities in society. This shift underscores the urgency for a thorough understanding and active steps to confront the ethical and societal consequences of AI bias, particularly within higher education contexts. Fu et al. [43] advocate for an integrative approach to ethical AI, which incorporates human psychological elements to better understand and address the development, function, and reduction of algorithmic bias. Such an approach aims to ensure the development and application of AI systems in a manner that is fair, responsible, and transparent, thereby fostering societal benefit.

Tackling the ethical challenges and societal repercussions of AI in higher education requires a diverse and comprehensive strategy. Technical strategies are crucial, focusing on the development of tools and frameworks dedicated to identifying and mitigating biases. Ntoutsi et al. [17] advocate for a shift from conventional AI algorithms that prioritize predictive performance, emphasizing the integration of ethical and legal principles into the design, training, and deployment of AI to achieve social benefits. This strategy necessitates addressing challenges in data-driven AI, particularly those arising from big data and advanced machine learning algorithms, and the potential for biased outcomes due to data collection and processing practices related to demographic factors like race and sex. Along the same lines, Dignum [44] explores the broader societal implications of AI, proposing a relational approach that considers the ethical, legal, societal, cultural, and environmental ramifications of AI technologies. This approach is based on the understanding that objective and rational reasoning does not always lead to the most appropriate outcomes, as the ‘right’ decision often varies depending on the specific context and dynamics at play. This perspective underscores the complexity of ethical decision-making in AI and highlights the need for nuanced and situationally aware approaches to AI development and implementation.

Woodgate and Ajmeri [45] introduce systematic methodologies aimed at consistently embedding normative ethical principles into the reasoning capacities of sociotechnical systems, where humans and technical agents collaborate. This approach equips practitioners with the tools to analytically and systematically address the complex social dilemmas that arise in such systems, aiming for resolutions that are satisfactory for all users involved. Guan et al. [46] followed these findings and also highlight technological uncertainty, incomplete data, and management errors as primary sources of ethical risks in AI decision-making. They propose comprehensive strategies for managing these ethical risks, offering perspectives from management, research, and development. This approach underscores the importance of a holistic governance strategy to oversee ethical challenges in AI decision-making processes.

Education and awareness in AI ethics are fundamental to addressing the ethical dilemmas and societal impacts of AI bias in higher education, as emphasized by Ali and Abdel-Haq [12]. They highlight the necessity of incorporating ethical reasoning into AI design and implementation, pointing towards the need for integrating AI ethics, law, and policy into educational curricula to prepare future AI developers and users. Köbis and Mehner [47] discuss ethical questions raised by AI-supported mentoring in higher education, emphasizing the need for a thorough understanding of ethical norms, guidelines, and unresolved issues in AI applications. Bendechache et al. [48] focus on engaging teenagers in workshops to reflect on ethical and privacy implications of AI, thus empowering them to evaluate the ethical aspects of AI in their lives. Benhayoun and Lang [49] identify gaps between academic training on AI and the requirements of the labor market, emphasizing the need for education to incorporate knowledge of ethical and regulatory dimensions of AI.

These studies and initiatives underscore the importance of a comprehensive strategy encompassing technical solutions, ethical guidelines, and an understanding of societal dynamics to address AI bias in higher education. They highlight that such an approach ensures that these systems contribute positively to society, promoting fairness and equity rather than perpetuating existing inequalities. The current state of the field indicates a growing recognition of the importance of ethical considerations in applications in higher education and an increasing effort to integrate ethics into AI education. However, challenges remain in ensuring the inclusivity and diversity of AI ethics education and effectively preparing students for the ethical dimensions of the technology in their future careers. The literature advocates for continued efforts to develop and implement comprehensive educational strategies that address these challenges and promote the ethical development and use of AI in higher education.

3. Methodology

3.1. Research design and data collection

This review article adopts a semi-systematic approach to rigorously examine the ethical challenges and biases of AI within the context of higher education. The primary research design involved a detailed literature review, focusing solely on published academic articles and reports rather than case studies, to ensure a comprehensive exploration of the topic. This clarification aligns with the abstract's revised focus solely on literature as the basis for data collection. The primary sources for this literature review were academic databases known for their extensive collections of peer-reviewed articles, including PubMed, IEEE Xplore, Google Scholar, and JSTOR. These platforms were selected for their relevance to artificial intelligence, ethics, and higher education fields.

The data collection process was meticulously structured; a combination of keywords such as "artificial intelligence in higher education", "ethics in AI", "bias in AI", "machine learning ethics", and "AI fairness" was used. These terms were carefully chosen to capture a wide array of relevant studies across a spectrum of AI's ethical implications in higher education. The inclusion criteria for selecting sources emphasized recent publications from 2010 onwards to reflect the most current trends and developments in the field. Publications that did not specifically address the ethical challenges or biases within AI in an educational context were excluded to maintain focus and relevance.

3.2. Data analysis

The collected literature underwent a thorough analysis where key information such as the authors' main arguments, methodologies, findings, and conclusions were extracted. This data was critically reviewed to identify prevailing themes, patterns, and gaps in the research concerning AI ethics in higher education. Each source was assessed for credibility, methodological rigor, and the significance of its contributions to the field, ensuring that the review rested on reliable and authoritative bases.

3.3. Presentation of findings

To enhance the clarity and impact of our findings, the synthesized data will be presented using various illustrative tools such as tables, diagrams, and figures. These visual representations will organize the extracted data effectively, making the complex information more accessible and easier to understand for readers. Tables will summarize the key findings and methodological approaches of the included studies, while figures will depict trends and relationships among the variables discussed in the literature.

3.4. Structuring the review

The findings from the literature review are structured into several distinct sections that logically flow from the general exploration of AI's implications to more specific discussions on ethical challenges and mitigation strategies. This structured approach ensures a coherent narrative that gradually builds upon initial concepts, leading to in-depth analyses and informed conclusions. The section ends with a discussion that ties together all insights gathered, addressing the initial research question while setting the stage for potential future research areas.

4. Results and recommendations

4.1. Overview of key findings

This review extensively explores the ethical challenges and biases presented by AI in the higher education sector, drawing upon a diverse array of studies spanning over a decade. Our semi-systematic approach led to the identification of significant trends in publication years, geographical distribution of research, key articles, and the ethical principles frequently discussed. Notably, the period from 2010 to 2022 marks an increase in scholarly attention towards AI ethics, reflecting the technology's advancing role in educational environments. The surge in publications around 2018 and 2020 coincides with pivotal advancements in AI capabilities and public awareness of its potential ethical implications.

The research reveals a global concern for AI ethics in higher education, with contributions from a wide range of countries, including but not limited to the United States, European nations, and several Asian and Australian regions. This geographical diversity underscores the universal relevance and urgency of addressing AI's ethical challenges across different educational and cultural contexts. Among the top-cited articles, works by Ntoutsi et al. [17] and Holmes et al. [50] stand out, offering comprehensive overviews and frameworks for understanding and mitigating bias in AI systems within the educational sector. These articles, among others, emphasize the necessity of embedding core ethical principles such as fairness, accountability, transparency, and privacy in the development and deployment of AI technologies.

4.2. Addressing bias through ethical frameworks

The commonly reported ethical principles identified in the literature—fairness, accountability, transparency, and privacy—serve as foundational elements guiding the discourse on ethical AI in education. These principles are crucial for developing AI systems that are not only technologically advanced but also equitable and beneficial to all stakeholders in the educational landscape. The inclusion of visual aids, such as charts and tables, in our review further facilitates an understanding of the research trends, geographical distribution, and thematic focuses within the field. The results of this review highlight the critical importance of addressing ethical challenges and biases in AI as it becomes increasingly integrated into higher education. The findings advocate for a multidisciplinary and globally inclusive approach to researching and implementing AI technologies, ensuring they align with ethical standards and contribute positively to educational outcomes. This comprehensive analysis sets the stage for future research directions and policy-making in the domain of ethical AI in education.

In the context of higher education, it is crucial to implement effective strategies for mitigating bias and promoting fairness in AI, with participatory design (PD) identified as a key methodology (see **Table 2**). Zytko et al. [51] underscore the importance of involving a diverse group of stakeholders in the AI design and development process through PD. This inclusive approach ensures that a wide range of perspectives and needs are considered, which is essential for addressing the negative societal impacts of AI and fostering a positive influence, particularly for vulnerable groups. Zytko and their colleagues highlight the growing application of PD in various domains, both in the private and public sectors. They emphasize the significant role of PD in reducing the detrimental effects of AI on society and enhancing its positive contributions, with a particular focus on supporting vulnerable populations. The panel organized by Zytko et al. [51] brought together experts in participatory design from different fields to discuss the practical and meaningful application of PD methods in AI systems, exploring the potential opportunities and challenges.

Table 2. Recommendations to address bias in AI in higher education.

Area of consideration	**Description**
Ethical frameworks for AI	Development of robust ethical frameworks that guide the entire lifecycle of AI systems, prioritizing transparency, justice, fairness, and privacy.
Interdisciplinary research in AI ethics	Bringing together experts from ethics, sociology, psychology, computer science, etc., to facilitate a holistic understanding of biases in AI and develop comprehensive mitigation strategies.
Innovative methods to mitigate bias	Exploring new algorithms, tools, and frameworks for detecting and mitigating bias effectively.
Personalized AI learning systems	Creating AI tools that adapt to diverse learning styles while avoiding cultural or socioeconomic biases, thus providing dynamic and responsive learning environments.
AI governance and policy development	Concentrating on developing comprehensive policies and regulatory frameworks that govern the ethical use of AI in education.
AI ethics education	Integrating AI ethics into the curriculum of higher education, developing courses and training programs to equip future AI developers and users with necessary ethical competencies.
Exploring societal impact of AI	Delving into the broader societal implications of AI in higher education, focusing on how AI systems can potentially reinforce or mitigate class-based power differences and stereotypes.
Public engagement and AI literacy	Enhancing AI literacy among students, faculty, and the broader community to ensure all stakeholders understand the capabilities, limitations, and ethical considerations of AI.
AI for inclusivity and accessibility	Prioritizing the development of AI tools that enhance inclusivity and accessibility in higher education, ensuring systems are accessible to students with disabilities and those from diverse cultural and socioeconomic backgrounds.

4.3. Enhancing interdisciplinary research

Previously, Goodyear [52] had taken a novel approach to educational design, particularly networked learning, as a context to outline patterns based approaches to educational design. The paper revisits the original conceptions of participatory design informed by the early work of Alexander (1977) on patterns and pattern languages. It connects the technicalities of design with the central place of values, which is crucial for encoding, sharing, and using knowledge for educational design. The paper emphasizes that a patterns-based approach is not only effective for sharing and re-using design experience but is also a powerful way of connecting educational values and vision to the tasks, tools, and resources offered to students.

Triantafyllakos et al. [53] introduced a framework for creating collaborative design games used in participatory design sessions with students for developing educational applications. This framework, drawing inspiration from idea generation theory and the literature on design games, guides the creation of board games that enable students to articulate their needs, desires, and expectations for future educational software. It was tested in various design sessions, demonstrating its effectiveness in fostering rapid and extensive exploration of the design space, and in eliciting a wide range of needs and ideas from participants.

Building on this concept, Hossain and Ahmed [54] advocated for the application of participatory design (PD) in developing AI technologies. They proposed an innovative agile participatory design approach, tailored not only for the design of AI and data-driven technologies but also to address existing challenges in the application of PD in this context. Their approach emphasizes a participatory, data-centric methodology for AI ethics by design, enhancing and applying insights from the family of value-sensitive design methods. These discussions and findings highlight the pivotal role of participatory design in mitigating bias and promoting fairness in AI within the higher education sector. By involving diverse stakeholders and focusing on ethical design principles, PD emerges as a key strategy in developing AI systems that are equitable and responsive to the needs of all users.

Along with PD, value sensitive design (VSD) is an essential strategy for critically analyzing and integrating specific desired values into the development of new technologies, particularly AI systems in higher education. Jacobs et al. [55] provide a comprehensive overview of VSD, examining its contributions to understanding and addressing bias in computer systems. They outline the current debates on algorithmic bias and fairness in machine learning and discuss how these debates could benefit from VSD-derived insights and recommendations. By focusing on values such as transparency, justice, fairness, and privacy, VSD helps guide the design of technologies that are not only functional but also ethically and socially responsible.

Umbrello and Van de Poel [56] presents a compelling argument for the use of Value Sensitive Design (VSD) as a method to integrate common values into AI systems from the early stages of development through to completion. Umbrello utilizes a case study of the UK Select Committee on artificial intelligence to demonstrate that different stakeholder groups involved in AI design and use often share similar values. These shared values can be harnessed to strengthen efforts in design coordination. Umbrello's work underscores the effectiveness of VSD in extracting these shared values and establishing a framework for coordinating stakeholders. Additionally, Gerdes [57] introduces a participatory, data-centric approach to AI ethics, rooted in value-sensitive design methodologies. This approach focuses on addressing both epistemic and ethical issues that arise during the early stages of machine learning development. By doing so, it opens up avenues for AI design that are informed by ethical considerations, enhancing the potential for creating AI systems that are grounded in ethical principles from the outset.

4.4. Innovating bias mitigation methods

Expanding on his previous research, Steven Umbrello, in collaboration with Ibo van de Poel [58], discusses the unique challenges that AI, especially machine learning, presents to value sensitive design (VSD). They suggest a revised version of the VSD methodology, one that incorporates a well-established set of principles to serve as design norms. From these norms, more detailed design requirements can be developed. This approach is designed to guarantee that the outcomes of AI development are not only harmless but also positively contribute to the greater good. Moreover, they advocate for an expansion of the VSD process to cover the entire lifecycle of AI technology, ensuring that ethical considerations are integrated from inception through deployment and beyond. This comprehensive approach aims to address the specific complexities and demands of AI and machine learning within the framework of value-sensitive design. These discussions and findings underscore the critical importance of VSD as a strategy for mitigating bias and promoting fairness in AI within higher education.

As these studies have demonstrated, addressing bias in AI within higher education involves a multi-layered strategy incorporating both technical and conceptual approaches to ensure fairness and equity. Kasif [58] introduces the concept of an "intelligent system quotient" as a measure to reflect the societal impact of AI systems. This quotient suggests a structured approach to understanding and mitigating AI bias through a multi-tier architecture, offering a quantifiable method to assess and address bias in AI, datasets, and algorithms.

4.5. Policy development and AI governance

AI impact assessment, as introduced by Nitta et al. [59] and Metcalf et al. [60], evaluates the potential impact of AI systems on society. Modeled after impact assessments in other fields, it scrutinizes the ethical implications of systems based on AI ethics guidelines and identifies ethical risks, providing a structured framework for accountability and responsibility in deployment. Along the same lines, Beutel [61] had previously proposed a set of best practice guidelines for selecting fairness metrics in AI models, helping practitioners avoid misjudging their AI models as fair and identifying conditions where certain fairness metrics may fail. The approach is crucial for developing more reliable and fair AI systems.

Finally, Broder and Berton [62] discuss the use of pre-processing algorithms for mitigating bias in machine learning models, as provided by IBM's AI fairness 360 framework. This technical approach is significant in analyzing and comparing the behavior of different learning algorithms when trained with biased data, marking a substantial advancement in the technical mitigation of bias in AI. Also, Han [63] proposes an augmented discriminator for adversarial training in natural language processing. The novel method, which utilizes the target class to create more nuanced features and explicitly model equal opportunity, has shown notable improvements in balancing performance and fairness. Aswell, it is particularly promising for addressing bias in natural language processing applications.

In analyzing the findings from the review in relation to the methods and results, it is evident that the strategies identified form a comprehensive framework for addressing bias in AI within higher education. The methodology of this paper, being semi-systematic, allowed for the inclusion of a diverse range of studies, providing a holistic view of the current state of AI in education, particularly focusing on ethical challenges and biases. The intelligent system quotient, as discussed in the reviewed literature, emerges as a pivotal tool for assessing the societal impact of AI systems. This measure, reflecting our methodological focus on recent and impactful studies, underscores the growing emphasis on quantifiable assessments in AI ethics. The AI Impact Assessment, another key strategy identified, aligns with the methodological criteria of relevance to current educational practices. This strategy, recognized in recent studies, highlights the necessity of evaluating AI systems for their potential societal impacts, an aspect that is increasingly being scrutinized in higher education.

The review's findings also emphasized specific technical and methodological strategies for mitigating bias in AI systems. These strategies, identified through a systematic analysis of current literature, resonate with the increasing call for more sophisticated and nuanced approaches to AI development in higher education settings. The emphasis on fairness and inclusivity in these approaches is reflective of the growing awareness and concern over AI biases, which was a key focus of the review process. Collectively, these strategies demonstrate the multi-faceted nature of addressing AI bias, aligning with the methods and results of the review. The emphasis on a combination of quantitative measures, impact assessments, and technical strategies highlights a comprehensive approach to ensuring AI technologies contribute positively and equitably to the educational landscape. This approach, informed by a thorough analysis of current and relevant literature, supports the creation of an inclusive and equitable educational environment, addressing the ethical challenges posed by AI in higher education.

5. Discussion

5.1. Analyzing ethical frameworks and interdisciplinary approaches

The findings of this review reveal a crucial dependence on robust ethical frameworks to guide AI applications throughout their lifecycle. As Gupta et al. [64] and Patel [65] highlighted, these frameworks need to emphasize transparency, justice, fairness, and privacy, adapting continually to the rapid advancements in AI technology. The interdisciplinary approach, which brings together experts from various fields, is not merely beneficial but necessary for a comprehensive understanding of biases in AI. This discussion underscores the importance of such frameworks in ensuring that AI technologies are designed and implemented with ethical considerations at their core, which supports the findings but also raises questions about the feasibility and scalability of such comprehensive approaches in diverse educational environments.

5.2. Effectiveness of technical methods in bias mitigation

Innovative technical methods for detecting and mitigating bias, as discussed by Schwartz et al. [29], are crucial in minimizing the unethical impacts of AI. However, the effectiveness of these methods needs continuous evaluation against empirical metrics, as suggested by Papamitsiou and Economides [66]. While these methods hold promise, the discussion should also consider the complexities and limitations inherent in developing tools that can universally identify and correct biases across different AI systems and contexts.

5.3. Governance, policy development, and AI ethics education

The role of governance and comprehensive policies in ethical AI usage, as proposed by Gellai [67], is another significant aspect that emerges from the findings. The necessity for embedding AI ethics into the curriculum to prepare future developers and users is crucial and reflects a proactive approach to ethical AI. However, the discussion should critically evaluate how existing educational policies accommodate these changes and the challenges in policy implementation across institutions with varying resources and priorities.

5.4. Societal implications and public engagement

The broader societal implications of AI in higher education, particularly its potential to reinforce or mitigate class-based power differences as discussed by Nelson and Zippel [68], raise essential considerations for equity in AI applications. While the findings advocate for increased public engagement and enhanced AI literacy, the discussion should critically assess the current state of public understanding and the practical steps needed to achieve widespread literacy and engagement.

5.5. Accessibility and inclusiveness of AI tools

The need for AI tools that enhance inclusivity and accessibility, highlighted by Mohammed and Nell'Watson [69], is well supported by the findings. However, the discussion should explore the real-world challenges in developing such tools, including technological limitations, funding constraints, and the potential for unintended exclusionary effects. This would involve examining how well these tools have been integrated into existing educational frameworks and their actual impact on students from diverse backgrounds.

5.6. Empirical support for proposed strategies

The suggestions for future research include using quantitative methods to assess the effectiveness of AI applications. While this approach is promising, the discussion must critically address the complexities involved in setting up such studies, the variability in data quality, and the interpretative challenges that may arise.

6. Conclusion

As higher education institutions increasingly adopt artificial intelligence (AI) and machine learning (ML), it is critical to navigate this technological evolution with a strong commitment to ethics. AI's potential to transform educational paradigms—from personalized learning environments to sophisticated predictive analytics in student support—offers significant opportunities for enhancing both learning experiences and operational efficiencies. However, these advancements come with the responsibility to address the ethical challenges they pose, particularly the potential biases that can undermine fairness, equity, and inclusivity in educational outcomes.

To ensure the ethical integration of AI in higher education, a multifaceted strategy is essential. This strategy must extend beyond the development of advanced technical solutions for bias detection and mitigation. There is a need for comprehensive ethical frameworks and governance policies that are continuously updated to reflect the evolving nature of AI technologies. Such frameworks should emphasize the creation of diverse and representative datasets, foster interdisciplinary collaboration, and adhere to inclusive design principles. These elements are critical in developing AI systems that are not only effective but also equitable and capable of serving the diverse needs of the student population.

Looking forward, the successful integration of AI into higher education depends on the concerted effort to weave ethical considerations into every facet of AI development and deployment. This includes enhancing the governance of AI technologies, promoting education on AI ethics, and actively engaging the public in discussions about AI's role in society. By investing in these areas, educational institutions can prepare a well-informed community of educators, students, and policymakers who are equipped to use AI responsibly and effectively.

As we stand on the brink of significant transformations in the educational sector driven by AI, it is our collective duty to guide the development and application of these technologies in ways that enhance educational practices and uphold core values of fairness, transparency, and inclusivity. This approach will ensure that AI in higher education not only fosters intellectual growth but also contributes to a culture of ethical awareness and social responsibility. In conclusion, by maintaining ethical vigilance and proactively addressing the challenges posed by AI, the academic community can harness its full potential to advance educational equity and excellence for all students.

Author contributions: Conceptualization, EB; methodology, EB; validation, EB; investigation, JH; writing—original draft preparation, JH; writing—review and editing, JH; visualization, JH. All authors have read and agreed to the published version of the manuscript.

Conflict of interest: The authors declare no conflict of interest.

References

1. Bisen IE, Arslan EA, Yildirim K, et al. Artificial Intelligence and machine learning in higher education. In: Machine Learning Approaches for Improvising Modern Learning Systems. IGI Global; 2021. pp. 1-17.

2. Rahiman HU, Kodikal R. Revolutionizing education: Artificial intelligence empowered learning in higher education. Cogent Education. 2023; 11(1). doi: 10.1080/2331186x.2023.2293431
3. Yildirim Y, Arslan EA, Yildirim K, et al. Reimagining Education with Artificial Intelligence. Eurasian Journal of Higher Education. 2021; 2(4): 32-46. doi: 10.31039/ejohe.2021.4.52
4. Alqahtani T, Badreldin HA, Alrashed M, et al. The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. Research in Social and Administrative Pharmacy. 2023; 19(8): 1236-1242. doi: 10.1016/j.sapharm.2023.05.016
5. Kuleto V, Ilić M, Dumangiu M, et al. Exploring Opportunities and Challenges of Artificial Intelligence and Machine Learning in Higher Education Institutions. Sustainability. 2021; 13(18): 10424. doi: 10.3390/su131810424
6. Alam A. Harnessing the Power of AI to Create Intelligent Tutoring Systems for Enhanced Classroom Experience and Improved Learning Outcomes. In: Intelligent Communication Technologies and Virtual Mobile Networks. Singapore: Springer Nature Singapore; 2023. pp. 571-591.
7. Pedró F. Applications of Artificial Intelligence to higher education: possibilities, evidence, and challenges. IUL Research. 2020; 1(1): 61-76. doi: 10.57568/iulres.v1i1.43
8. Sajja R, Sermet Y, Cikmaz M, et al. Artificial Intelligence-Enabled Intelligent Assistant for Personalized and Adaptive Learning in Higher Education. arXiv. 2023; arXiv:2309.10892.
9. Xu W, Meng J, Raja SKS, et al. Artificial intelligence in constructing personalized and accurate feedback systems for students. International Journal of Modeling, Simulation, and Scientific Computing. 2023; 14(01): 2341001.
10. Miao F, Holmes W, Huang R, et al. AI and education: A guidance for policymakers. UNESCO Publishing; 2021.
11. Yang C, Huan S, Yang Y. A Practical Teaching Mode for Colleges Supported by Artificial Intelligence. International Journal of Emerging Technologies in Learning (iJET). 2020; 15(17): 195. doi: 10.3991/ijet.v15i17.16737
12. Ali M, Abdel-Haq MK. Bibliographical analysis of artificial intelligence learning in Higher Education: is the role of the human educator and educated a thing of the past? In: Fostering Communication and Learning with Underutilized Technologies in Higher Education. IGI Global; 2021. pp. 36-52.
13. Tiwari CK, Bhat MohdA, Khan ST, et al. What drives students toward ChatGPT? An investigation of the factors influencing adoption and usage of ChatGPT. Interactive Technology and Smart Education. Published online August 29, 2023. doi: 10.1108/itse-04-2023-0061
14. Popenici SAD, Kerr S. Exploring the impact of artificial intelligence on teaching and learning in higher education. Research and Practice in Technology Enhanced Learning. 2017; 12(1). doi: 10.1186/s41039-017-0062-8
15. Zawacki-Richter O, Marín VI, Bond M, et al. Systematic review of research on artificial intelligence applications in higher education—where are the educators? International Journal of Educational Technology in Higher Education. 2019; 16(1): 1-27.
16. Maphosa V, Maphosa M. Artificial intelligence in higher education: a bibliometric analysis and topic modeling approach. Applied Artificial Intelligence. 2023; 37(1). doi: 10.1080/08839514.2023.2261730
17. Ntoutsi E, Fafalios P, Gadiraju U, et al. Bias in data-driven artificial intelligence systems—An introductory survey. WIREs Data Mining and Knowledge Discovery. 2020; 10(3). doi: 10.1002/widm.1356
18. Hall P, Ellis D. A systematic review of socio-technical gender bias in AI algorithms. Online Information Review. 2023; 47(7): 1264-1279. doi: 10.1108/oir-08-2021-0452
19. Kordzadeh N, Ghasemaghaei M. Algorithmic bias: review, synthesis, and future research directions. European Journal of Information Systems. 2021; 31(3): 388-409. doi: 10.1080/0960085x.2021.1927212
20. Leavy S. Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In: Proceedings of the 1st international workshop on gender equality in software engineering; 2018. pp. 14-16.
21. Yapo A, Weiss J. Ethical Implications of Bias in Machine Learning. In: Proceedings of the Annual Hawaii International Conference on System Sciences; 2018. doi: 10.24251/hicss.2018.668
22. Mehrabi N, Morstatter F, Saxena N, et al. A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys. 2021; 54(6): 1-35. doi: 10.1145/3457607
23. Zhou KZ, Cao J, Yuan X, et al. "I'm Not Confident in Debiasing AI Systems Since I Know Too Little": Teaching AI Creators About Gender Bias Through Hands-on Tutorials. arXiv. 2023; arXiv:2309.08121.
24. Roshanaei M. Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education. Education and Information Technologies. 2024; 1-26.

25. Broussard M. More than a glitch: Confronting race, gender, and ability bias in tech. MIT Press; 2023.
26. Jackson S, Panteli N. Trust or mistrust in algorithmic grading? An embedded agency perspective. International Journal of Information Management. 2023; 69: 102555. doi: 10.1016/j.ijinfomgt.2022.102555
27. Slimi Z, Carballido BV. Navigating the Ethical Challenges of Artificial Intelligence in Higher Education: An Analysis of Seven Global AI Ethics Policies. TEM Journal. 2023; 12(2): 590-602. doi: 10.18421/tem122-02
28. Le-Nguyen HT, Tran TT. Generative AI in Terms of Its Ethical Problems for Both Teachers and Learners: Striking a Balance. In: Generative AI in Teaching and Learning. IGI Global; 2023. pp. 144-173.
29. Schwartz R, Vassilev A, Greene K, et al. Towards a standard for identifying and managing bias in artificial intelligence. NIST special publication; 2022.
30. Hersh RH, Merrow J. Declining by degrees: Higher education at risk. St. Martin's Press; 2015.
31. Mashhadi A, Zolyomi A, Quedado J. A Case Study of Integrating Fairness Visualization Tools in Machine Learning Education. In: Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts. doi: 10.1145/3491101.3503568
32. Islam SR, Russell I, Eberle W, et al. Incorporating the Concepts of Fairness and Bias into an Undergraduate Computer Science Course to Promote Fair Automated Decision Systems. In: Proceedings of the 53rd ACM Technical Symposium on Computer Science Education V2. pp. 1075-1075. doi: 10.1145/3478432.3499043
33. Borenstein J, Howard A. Emerging challenges in AI and the need for AI ethics education. AI and Ethics. 2020; 1(1): 61-65. doi: 10.1007/s43681-020-00002-7
34. Aleven V, Baraniuk R, Brunskill E, et al. Towards the Future of AI-Augmented Human Tutoring in Math Learning. In: International Conference on Artificial Intelligence in Education. Cham: Springer Nature Switzerland; 2023. pp. 26-31.
35. Giani MS, Walling D. Will I Get In? Using Predictive Analytics to Develop Student-Facing Tools to Estimate University Admissions Decisions. Journal of College Access. 2020; 5(1): 4.
36. Smithers L. Predictive analytics and the creation of the permanent present. Learning, Media and Technology. 2022; 48(1): 109-121. doi: 10.1080/17439884.2022.2036757
37. Umer R, Susnjak T, Mathrani A, et al. Current stance on predictive analytics in higher education: opportunities, challenges and future directions. Interactive Learning Environments. 2021; 31(6): 3503-3528. doi: 10.1080/10494820.2021.1933542
38. Fuller K, Lupton-Smith C, Hubal R, et al. Automated Analysis of Preceptor Comments: A Pilot Study Using Sentiment Analysis to Identify Potential Student Issues in Experiential Education. American Journal of Pharmaceutical Education. 2023; 87(9): 100005. doi: 10.1016/j.ajpe.2023.02.005
39. Beal CR. AnimalWatch: An intelligent tutoring system for algebra readiness. In: International handbook of metacognition and learning technologies. Springer New York; 2013. pp. 337-348.
40. Chu CH, Nyrup R, Leslie K, et al. Digital Ageism: Challenges and Opportunities in Artificial Intelligence for Older Adults. Meeks S, ed. The Gerontologist. 2022; 62(7): 947-955. doi: 10.1093/geront/gnab167
41. Tomalin M, Byrne B, Concannon S, et al. The practical ethics of bias reduction in machine translation: Why domain adaptation is better than data debiasing. Ethics and Information Technology. 2021; 1-15.
42. Li C, Xing W. Revealing factors influencing students' perceived fairness: A case with a predictive system for math learning. In: Proceedings of the Ninth ACM Conference on Learning@ Scale; 2022. pp. 409-412.
43. Fu R, Huang Y, Singh PV. AI and algorithmic bias: Source, detection, mitigation and implications. In: Detection, Mitigation and Implications. 2020.
44. Dignum V. Relational artificial intelligence. arXiv. 2022; arXiv:2202.07446.
45. Woodgate J, Ajmeri N. Macro ethics for governing equitable sociotechnical systems. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems; 2022. pp. 1824-1828.
46. Guan H, Dong L, Zhao A. Ethical Risk Factors and Mechanisms in Artificial Intelligence Decision Making. Behavioral Sciences. 2022; 12(9): 343. doi: 10.3390/bs12090343
47. Köbis L, Mehner C. Ethical questions raised by AI-supported mentoring in higher education. Frontiers in Artificial Intelligence. 2021; 4: 624050.
48. Bendechache M, Tal I, Wall P, et al. AI in My Life: AI, Ethics & Privacy Workshops for 15-16-Year-Olds. In: Proceedings of the Companion Publication of the 13th ACM Web Science Conference; 2021. pp. 34-39.
49. Benhayoun L, Lang D. Does higher education properly prepare graduates for the growing artificial intelligence market? Gaps' identification using text mining. Human Systems Management. 2021; 40(5): 639-651. doi: 10.3233/hsm-211179

50. Holmes W, Porayska-Pomsta K, Holstein K, et al. Ethics of AI in Education: Towards a Community-Wide Framework. International Journal of Artificial Intelligence in Education. 2021; 32(3): 504-526. doi: 10.1007/s40593-021-00239-1
51. Zytko D, J. Wisniewski P, Guha S, P. S. Baumer E, Lee MK. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In: Proceedings of the CHI Conference on Human Factors in Computing Systems Extended Abstracts. doi: 10.1145/3491101.3516506
52. Goodyear P. Educational design and networked learning: Patterns, pattern languages and design practice. Australasian Journal of Educational Technology. 2005; 21(1). doi: 10.14742/ajet.1344
53. Triantafyllakos G, Palaigeorgiou G, Tsoukalas IA. Designing educational software with students through collaborative design games: The We! Design & Play framework. Computers & Education. 2011; 56(1): 227-242.
54. Hossain S, Ahmed SI. Towards a New Participatory Approach for Designing Artificial Intelligence and Data-Driven Technologies. arXiv. 2021; arXiv:2104.04072.
55. Jacobs M, Kurtz C, Simon J, et al. Value Sensitive Design and power in socio-technical ecosystems. Internet Policy Review. 2021; 10(3). doi: 10.14763/2021.3.1580
56. Umbrello S, Van de Poel I. Mapping value sensitive design onto AI for social good principles. AI and Ethics. 2021; 1(3): 283-296.
57. Gerdes A. A participatory data-centric approach to AI Ethics by Design. Applied Artificial Intelligence. 2021; 36(1). doi: 10.1080/08839514.2021.2009222
58. Kasif S. Artificial Tikkun Olam: AI Can Be Our Best Friend in Building an Open Human-Computer Society. arXiv. 2020; arXiv:2010.12015.
59. Nitta I, Ohashi K, Shiga S, et al. AI Ethics Impact Assessment based on Requirement Engineering. In: Proceedings of the 2022 IEEE 30th International Requirements Engineering Conference Workshops (REW). doi: 10.1109/rew56159.2022.00037
60. Metcalf J, Moss E, Watkins EA, et al. Algorithmic Impact Assessments and Accountability. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. doi: 10.1145/3442188.3445935
61. Beutel A, Chen J, Doshi T, et al. Putting fairness principles into practice: Challenges, metrics, and improvements. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp. 453-459.
62. Broder RS, Berton L. Performance analysis of machine learning algorithms trained on biased data. In: Anais do XVIII Encontro Nacional de Inteligência Artificial e Computacional. SBC; 2021. pp. 548-558.
63. Han X, Shen A, Li Y, et al. FairLib: A Unified Framework for Assessing and Improving Fairness. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. doi: 10.18653/v1/2022.emnlp-demos.7
64. Gupta A, Wright C, Ganapini MB, et al. The State of AI Ethics Report. arXiv. 2021; arXiv:2108.03929.
65. Patel K. Ethical Reflections on Data-Centric AI: Balancing Benefits and Risks. International Journal of Artificial Intelligence Research and Development. 2024; 2(1): 1-17.
66. Papamitsiou Z, Economides AA. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. Journal of Educational Technology & Society. 2014; 17(4): 49-64.
67. Gellai DB. Enterprising Academics: Heterarchical Policy Networks for Artificial Intelligence in British Higher Education. ECNU Review of Education. 2022; 6(4): 568-596. doi: 10.1177/20965311221143798
68. Nelson LK, Zippel K. From Theory to Practice and Back: How the Concept of Implicit Bias was Implemented in Academe, and What this Means for Gender Theories of Organizational Change. Gender & Society. 2021; 35(3): 330-357. doi: 10.1177/08912432211000335
69. Mohammed PS, Nell'Watson E. Towards inclusive education in the age of artificial intelligence: Perspectives, challenges, and opportunities. Artificial Intelligence and Inclusive Education: Speculative futures and emerging practices. 2019; 17-37.