Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 15, No. 3, 2024

The impact of AI bias on social justice: challenges and solutions

Jiali Ma*

College of Computer Science and Technology, Jilin University, Changchun, Jilin,130012, China.

* **Corresponding author Email:** mjl20030423@163.com

Abstract: This paper analyzes the wide-ranging impact of artificial intelligence (AI) bias on social equity, especially in key areas such as healthcare, education, and employment. With the rapid development of AI technology, its potential bias problems have been gradually exposed and attracted widespread attention. To this end, this paper not only explores the manifestations and causes of AI biases, but also proposes specific mitigation strategies, such as through improving data diversity, enhancing algorithm transparency, promoting interdisciplinary cooperation, and establishing a sound regulatory framework, with a view to reducing the negative impact of AI biases on social equity and promoting the fair application of AI technologies in various fields of society.

Keywords: Artificial Intelligence Bias, Social Justice, Data Bias, Human Bias, Algorithmic Bias.

1. Introduction

Artificial Intelligence (AI) has demonstrated a strong impact in recent years in a variety of fields such as healthcare, finance, and education. It has dramatically revolutionized traditional work patterns by analyzing large amounts of data, identifying complex patterns and making efficient predictions. This technological advancement not only improves the efficiency of the industry, but also promotes innovation. For example, AI can help doctors diagnose diseases more accurately in the medical field, optimize investment strategies in the financial field, and personalize the learning experience in the education field. However, with the wide application of AI technology, the problem of bias in AI systems is gradually exposed.

These biases have permeated our technological development and daily lives, and have challenged social justice and ethics. Addressing AI biases has become particularly urgent against the backdrop of the rapid development of AI technology. Unfair behavior of AI systems can have far-reaching negative impacts on society, especially in areas involving sensitive decision-making, such as the judicial system, healthcare, and education. To ensure that AI technologies can bring relative fairness to all aspects of society and maximize the benefits they bring, it is crucial to study and address the issue of AI bias.

The main contribution of this paper is to provide a comprehensive perspective on AI bias for both academia and industry, helping researchers and practitioners to better understand and respond to the issue, thereby ensuring that AI technologies are more fair, transparent, and explainable in future applications.

2. What is AI bias

Prejudice is a feature of human life that is intertwined with many different names and labels, or used interchangeably Stereotypes, prejudices, implicit or subconsciously held beliefs, or closed-mindedness" [2]. Social psychologists define "prejudice" as an unfair attitude towards a social group and its members, a prior or pre-existing judgment [3].

Research suggests that AI's predictions may systematically deviate from the truth, generating results that favor or disfavor certain groups. Such biases may stem from design flaws, skewed training data, confounding algorithmic models, or a lack of algorithmic computational power, and these biases can lead to harmful societal consequences and exacerbate widespread concerns about AI. Defining and identifying AI bias requires care to avoid misinterpreting or over-attributing AI, which can lead to unnecessary fear or misuse of the technology.

3. AI bias in education

3.1. Application status of intelligent education system

Currently, there are heated discussions about Artificial Intelligence in Education (AIED). Whether as a tool to assist teaching, campus management, or an important part of subject education, AI has injected a new impetus into the development of education.

On the one hand, AIED and the corresponding innovative education models can promote the sharing of quality resources, bridge the education gap and promote the process of education equity; on the other hand, however, the prejudices brought about by the distribution of the social power structure may be consolidated and exacerbated by the pursuit of adaptive teaching and personalized learning. Now that we are standing at this crossroads, we should clarify the relationship between inclusion and equity, and adaptation and individuality in education, and fully understand the significance of studying the bias in AIED.

3.2. Use cases of AI in education and the formation and impact of bias

3.2.1. Developer cognitive bias affects educational equity

The Teacher Intelligence Assessment Model (TIA) has caused a storm in the United States. The model compares what a student might achieve with a moderately competent teacher with what the student would actually achieve with a particular teacher. The former is the predicted grade given by the system and the latter is the actual grade of the student. The comparison automatically labels the teacher as "effective," "efficient," "developing," or "ineffective. The system automatically labels teachers as "effective," "efficient," "developing," and "ineffective. These labels affect teachers' reputations, salaries, and retention. One of the major debates about such intelligent assessment systems is whether teacher ratings should be based on the progress of students' test scores. If a student already scored high on the last test, there's little room for improvement on the next one. Although Lederman's students did not make significant progress in 2014, they did well in 2013, so her teaching cannot be arbitrarily judged "ineffective. The technology developer's cognitive bias of "causality between teacher effort and student progress" and the overly reliance on student progress in evaluating teachers' work nearly led to the dismissal of an excellent teacher. [11]

3.2.2. Discrimination due to unequal distribution of resources in the age of AI-enhanced intelligence

In an intelligent, student-centered learning environment, bias can lead to the exclusion of specific groups of students, exacerbating discrimination and segregation between groups in terms of gender, culture, geography, level of knowledge, and economic level of the family. A case in point is the Institute for Innovation in Public School Choice (IIPSC) in the United States, which has developed a high school admissions recommendation system that collects and analyzes three types of data: what schools parents want their children to attend, the number of seats available in each school and grade, and the school's admissions requirements and regulations. By collecting and analyzing three types of data: what schools parents want their children to attend, the number of seats available at each school and grade level, and the school's admission requirements and regulations, the system recommends high schools for students in New York, Boston, and Denver [13]. It has been found that this system forces parents and students to make choices with opaque and incomplete information, and does not improve the situation of disadvantaged groups of students, who are mostly recommended to out-of-town or inferior high schools [11]. Moreover, due to the "black box" nature of the algorithms, it is difficult for students, teachers, parents, and administrators who lack specialized technical knowledge to review the output of the system.

4. AI bias in healthcare

4.1. Development and Application of Medical AI Systems

Medical artificial intelligence (AI) has become increasingly valuable for disease diagnosis, monitoring and treatment, and plays an important role in various aspects such as risk factor identification, health resource allocation and precision medical intervention [14]. AI can provide more accurate medical diagnosis and treatment plans by analyzing data such as medical images, biomarkers, and medical records, reducing non-subjective errors and alerting doctors to possible cognitive biases, reducing medical misdiagnosis and omission, and improving the quality and fairness of healthcare services, thus realizing fairer health decision-making; it can also expand medical resources to remote areas and socially under-resourced areas by means of telemedicine and smart medical devices, enabling low-income groups and people in underdeveloped areas to access richer and more specialized healthcare resources and promoting public health [15]. Through telemedicine and smart medical devices, medical resources can also be expanded to remote areas and areas with scarce social resources, so that low-income groups and people in underdeveloped areas can have access to richer and more specialized healthcare resources, reducing the urban-rural healthcare gap and promoting public health [15].

4.2. Examples of AI Bias in Healthcare

4.2.1. Bias due to lack of transparency in the AI system

Lack of transparency in the output application of AI The determination of algorithmic bias consists of two conditions: one is that the AI treats disadvantaged groups differently, and the other is that the AI has an actual adverse effect on disadvantaged groups compared to advantaged groups. However, the "judgmental relevance" and "black box" nature of AI technology has led to a lack of transparency in the construction and application of AI by healthcare workers, which makes it difficult to assess the impact of human-selected content on the logic and output of AI, thus exacerbating algorithmic bias. The problem is that it is difficult to assess the impact of human-selected content on AI logic and output, thus exacerbating algorithmic bias. [16]

Second, AI technology is often viewed as a "black box", i.e., highly complex AI computational processes are often beyond human cognition and understanding, making it difficult for healthcare workers to clarify and explain the reasons for AI outputs [17]. Ultimately, clinical decision-making based on AI bias outputs will lead to health equity issues and exacerbate hidden problems of data generation and collection in the course of practice. Ultimately, clinical decisions based on AI bias will induce health equity issues and exacerbate the implicit bias of data generation and collection in the process of practical application, creating a vicious cycle of algorithmic bias [18].

4.2.2. Developer bias in the healthcare system

Algorithm designers, possibly unconsciously or by specific design in a trade-off of interests, embed social biases into the system design, producing outputs that are unfavorable to disadvantaged groups and biasing the algorithm. In December 2020, frontline medical staff at Stanford University Medical Center (Stanford Health Care) protested against the hospital because they had not received the first batch of the new Champion vaccine. Hospital administrators blamed an algorithm for the incident, arguing that the algorithm determined the order in which hospital employees were vaccinated. However, experts analyzed the algorithm in detail and found that the real mistake was made by the people who designed the algorithm, which did not determine the order of vaccinations based on the level of exposure of employees to the virus, but rather simply ranked them according to age. This case illustrates the fact that algorithmic decision-making is not always fair, and that algorithms are often easy scapegoats for decision-makers when things go wrong [19].

4.2.3. Healthcare workers' and patients' biases against AI promote prejudice

Patients' own socioeconomic status, race, ethnicity, and geography lead to unequal access to resources for doctors, hospitals, and other healthcare services, and this inequality extends to the availability of healthcare AI technology, an issue for all health systems around the world. Second, patient trust in AI varies. Some patients may know little about AI technology or may even be skeptical of their doctors or providers, and these patients may be more distrustful of their doctors' or providers' use of AI, which leads to variability in the prevalence of AI, which in turn affects the efficiency and impact of healthcare visits.

While traditional clinical decision support systems (CDSS) rely on specialized knowledge and guidelines for decision making, AI-based models rely more on statistical correlations. However, AI algorithms often lack transparency or interpretability in translating data into model outputs, and such statistical correlations may involve variables or interactions that physicians do not expect. As the medical community focuses on expert experience and specialized knowledge, clinicians' recommendations for AI tools and perceptions of transparency and interpretability around AI and its use of data may influence their willingness to use AI tools [20].

4.3. Solutions and Mitigation Strategies

4.3.1. Minimizing bias in intelligent medicine due to differences in individual circumstances

Social determinants of health (SDOH) Refers to the full range of socio-environmental characteristics of people from birth to aging, which include income, education, water, food and sanitation, living conditions, and access to health care. Medical interventions do not address the causes of human illness, and SDOH plays a critical role in disease risk and regression. The development of healthcare AI should integrate complete contextual information about a patient's health status, and we can make the model more comprehensive by collecting relevant social indicators or health influencers. However, the lack of uniform standards to capture the diversity of different patient groups and their unique social factors may also lead to bias, which requires the establishment of a data standard that includes health-related factors such as race, ethnicity, gender, disability status, work, preference, life experience, and so on. Studies have aggregated tests of homogeneity of medical AI for groups such as European, African-American, Asian, Latino, and Native American, demonstrating that factors such as different ethnicities and social exposures can lead to different outcome risk attributes, and therefore, explicitly collecting detailed demographic data criteria can optimize medical AI for more accurate decision making. [21]

4.3.2. Achieving greater transparency in AI systems

Now that AI algorithms are becoming more and more complex, we should appropriately increase the transparency and traceability of algorithms. Algorithm transparency includes the obligation to inform, report parameters to the competent authorities, disclose parameters to the society, archive data and disclose source code, etc. The right to interpret algorithms can be regarded as a concrete embodiment of the principle of algorithm transparency. Laws related to algorithmic transparency should be accelerated to strengthen algorithmic governance, and some scholars have suggested that disclosure of algorithmic source code should be set as a legal obligation for companies in order to improve the medical AI regulatory system [46]. Finally, healthcare workers need to strengthen the transparency of AI output application, deepen their understanding of clinical tasks and AI technology, . It is also necessary to consider how to identify and manage unfairness and artificial bias in clinical practice to maintain patient health equity. [22]

5. Conclusion

This paper explores the multifaceted issue of AI bias and its impact on various fields such as education, healthcare, and employment. Through detailed analysis, we find that AI bias can come from multiple sources, such as biased data, algorithm design flaws, and inherent human biases. The possibility of biased outcomes from artificial intelligence (AI) is no longer a new topic. Since AI algorithms are derived from human design choices, which are often difficult to completely free from values, bias is inevitable. One of the key findings is that AI systems, while powerful and transformative, are not immune to the biases embedded in the data and decisions on which they rely. For example, in healthcare, biased AI algorithms can lead to unequal treatment and misdiagnosis, especially among marginalized groups. In education, AI-powered systems may reinforce stereotypes or impede the academic progress of certain groups of students. These examples demonstrate the urgent need for more robust mechanisms to identify, mitigate and prevent AI bias.

References

[1] CALISKANA, BRYSON J J, NARAYANAN A. Semantics derived automatically from language corpora contain human-like biases[J]. Science,2017, 356(6334):183-186.

[2] Howard, A., & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. Science and Engineering Ethics, 24(5), 1521-1536.

[3] ZHANG Zhong-zhong,SONG Juan. Advances in prejudice research[J]. Psychological and Behavioral Research, 2007(2): 150-155.

[4] Del Balso, M. And Hermann, J. (2017). "Meet Michelangelo: Uber's Machine Learning Platform". Uber Engineering, 5 Sep 2017.

[5] Drew Roselli, Jeanna Matthews, and Nisha Talagala. 2019. Managing Bias in AI. In Companion Proceedings of the 2019 World Wide Web Conference, San Francisco, CA USA, May 2019 (WWW '19 Companion), 10 pages. Doe: 10.1145/3308560.3317590

[6] S. Townson and M. Zeltkevic, "Why AI bias may be easier to fix than humanity's," World Economic Forum Agenda, Jun. 2023. [Online]. Available: https://www.weforum.org/agenda/2023/06/why-ai-bias-may-be-easier-to-fix-than-humanity-s/. [Accessed: Sep. 01, 2024].

[7] ROSE A. Are face -detection cameras racist? [EB/OL]. (2010 -01 -22) [2021 -01 -26]. Http: //content.time.com/time/business/article/ 0,8599, 1954643, 00.html.

[8] Diakopoulos N. Algorithmicaccountability: Journalistic Investigation of Computational Power Structures [J]. Digitaljournalism, 2015,3(3):398-415.

[9] Yang Thanh Viet,Luo Xianjue. A preliminary study on the comprehensive management of algorithmic discrimination[J]. Science and Society, 2018,8(4):8-19

[10] World Educators Forum. Education 2030: incheon declaration and framework for action towards inclusive and equitable quality education and Inclusive and equitable quality education and lifelong learning for all[M]. Paris: United Nations Educational, Scientific and Cultural Organization, 2015.

[11] HARRIS E. Court vacates Long Island teacher's evaluation tied to test scores [EB/OL]. (2016-05-10) [2021-01-26]. Https: //www. Nytimes.com/2016/05/11/nyregion/court-vacates-

long-island-teachers-evaluation-tied-to-student-test-scores.html.

[12] ANGWIN J, Mattu S, LARSON J. The tiger mom tax: Asians are nearly twice as likely to get a higher price from Princeton review [EB/OL]. (2015-09-01) [2021-01-26]. Https://www.propublica.org/article/asians-nearly-twice-as-likely-to-get-higher-price-from Princeton-review

[13] HEROLD B. Custom software helps cities manage school choice [EB/OL]. (2013-11-03) [2021-01-26]. Https://www.edweek.org/ew/Articles/2013/12/04/13algorithm_ep.h33.html.

[14] MHASAWADEV, ZHAO Y, CHUNARA R. Machine learning and algorithmic fairness in public and population health[J]. Nat Mach intell, 2021, 3 (8): 659-666.

[15] ZHANG Yifan,ZHANG Zerui,DONG Jing,et al. Progress and challenges of medical artificial intelligence technology in the era of big model[J]. China Medical Equipment, 2024, 21(06): 189-194.

[16] Zhan Hao . Research on Algorithmic Discrimination in Data Mining in Big Data Era [D]. Changsha: Hunan Normal University, 2020.

[17] LEPRI B,OLIVER N,LETOUZÉ E, et al. Fair, transparent, and accountable algorithmic decision-making processes[J]. Philosophy & Technology,2018, 31: 611-627. DOI: 10.1007/s13347-017-0279-x.

[18] CHU C H, NYRUP R, LESLIE K, et al. Digital ageism: challenges and opportunities in artificial intelligence for older adults [J]. Gerontologist, 2022, 62(7): 947-955. DOI: 10.1093/geront/gnab167.

[19] J. Wang. From human bias to algorithmic bias: Can bias be eliminated[J]. Exploration and Controversy,.2021(3): 32-34

[20] FELDMAN R, ALDANA E, STEIN K. Artificial Intelligence in the Health Care Space: How We Can Trust What We Cannot Know[J]. Stanford Law And Policy Review, 2019, 30(2): 399.

[21] DANKWA-MULLAN I, WEERARATNE D. Artificial intelligence and machine learning technologies in cancer care: addressing disparities, bias, and data diversity[J]. Cancer Discov, 2022, 12(6): 1423-1427.

[22] Nature Machine Intelligence. Striving for health equity with machine learning [J]. Nature Machine Intelligence, 2021, 3: 653. DOI: 10.1038/s42256-021-00385-0.