Empowering Individuals: Cultivating Literacy to Mitigate AI Bias

Sungha Kang
University of Washington
Seattle, USA
sungha24@uw.edu

Abstract

Artificial intelligence (AI) has emerged as a powerful epistemic technology that is changing the way we collect, analyze, and interpret information to create knowledge. However, concerns are growing about the transparency and trustworthiness of AI systems. This position paper examines the discourse surrounding these important issues and highlights the importance of empowering individuals to engage critically with AI technologies. The paper discovers the various interventions currently being discussed in the community – Fair and ethical design and implementation of AI applications, Reducing bias from the algorithm, Ensuring diversity in AI development – and why fostering individual AI literacy is more primary and important than these technical and policy framework interventions. AI Literacy education with priority strategies of Critical and Ethical Thinking Skills and Accessible and Inclusive Learning will be the most powerful strategy in reducing the impacts and harms of AI bias.

Keywords: Artificial Intelligence (AI), Bias mitigation, Critical thinking, AI Literacy

ACM Reference Format:
Sungha Kang. 2024. Empowering Individuals: Cultivating Literacy to Mitigate AI Bias. In *Proceedings of Dr. Bill Howe (IMT589 Epistemological Foundations of AI).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Artificial intelligence (AI) stands as a transformative technology reshaping the landscape of human knowledge acquisition and decision-making processes. With its ability to process vast amounts of data, recognize patterns, and make predictions, AI has fostered innovation in various fields by providing insights and solutions previously thought unattainable. But amid the excitement surrounding the capabilities of AI, concerns are growing about its transparency, trustworthiness, and susceptibility to bias [21].

The importance of these concerns should never be overlooked. As AI systems become increasingly integrated into various aspects of society, their opacity and potential biases raise serious ethical and epistemological issues. There is ongoing discourse surrounding the ethical deployment of technology. The discussions start with "Can we trust the processes and results of AI algorithms?" followed by "Then how do we ensure transparency and trustworthiness in AI systems?" and "But, does ensuring transparency mean that AI algorithm results can be trusted?" [6, 18]

This position paper argues that individual empowerment is the most essential strategy for mitigating AI bias. It first explores the concept of AI bias and reviews real-world case studies highlighting its impact on AI systems. After looking at the various strategies discussed to mitigate AI bias, we will consider why empowering individuals, including literacy education, is the most important. Lastly, we will look in detail at the factors and potential strategies that must be considered important in this regard.

Central to the discussion is the role of self-directed learning in increasing individuals' AI capabilities. As users of AI technologies, individuals must have the knowledge and skills to engage critically with AI systems and discern between trustworthy insights and potentially biased results. Potential strategies and educational initiatives enable individuals to actively participate in shaping the ethical trajectory of AI development and deployment. Even if various intervention strategies are developed, we will always face bias and discrimination in technologies, including AI. Therefore, the most basic yet important intervention is to empower individuals.

By empowering individuals to utilize AI intelligently, we hope to contribute to the creation of an ethically responsible AI ecosystem. As intellectual leaders, they will pay greater attention to the ethical requirements of AI deployment, actively combat bias, and ensure that the promise of AI innovation is realized in a way that upholds social values and principles.

IMT589 Epistemological Foundations of AI, 2024, Seattle, WA

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

2 Understanding AI Bias

AI bias refers to systematic errors or inaccuracies in algorithms that disproportionately affect certain individuals or groups and lead to unfair outcomes based on characteristics such as race, gender, and socioeconomic status [7]. Bias can take many forms, including selection bias, where the training data used to develop AI models fails to represent the diversity of the population, and algorithmic bias, where the decision-making process of an AI system perpetuates existing social biases [12].

AI bias has been demonstrated in multiple contexts. Real-world examples discussed in past scholarships provide miserable illustrations of AI bias's impact on individuals and communities. Case studies from healthcare, employment, and criminal justice highlight how biased algorithms can lead to unequal outcomes [1–3]. For example, biased AI diagnostic tools in healthcare can misdiagnose certain demographic groups, leading to disparities in treatment and outcomes [11]. Likewise, a biased hiring algorithm can perpetuate systemic inequalities by favoring certain demographic groups over others [3].

In one experiment I conducted in April 2024, when ChatGPT4 was asked to draw a doctor and a nurse, it depicted a male (old) and a female (young) nurse. Upon questioning, it justified this depiction by suggesting that the doctor is typically a professional and older, hence the male portrayal. Conversely, when prompted for a young doctor and an old nurse, it depicted a male (young) doctor and a female (old) nurse. Even in this trivial case, we can find that AI has various biases related to gender, age, occupation, etc., and it is unimaginable to estimate what impact this will have on individuals and society.

The implications of AI bias go far beyond individual cases and encompass broader societal ramifications. Biased algorithms reinforce existing inequalities and undermine trust in AI technologies and institutions [5]. Moreover, they can exacerbate social divisions and contribute to systemic injustice [10]. As AI continues to permeate various aspects of society, it is essential to address bias to ensure fairness, equity, and social justice.

3 Interventions for AI Bias

The research community is aware of the problem of AI bias and is continuously working on research to mitigate it. However, as seen from the preceding definition and several examples, mitigating AI bias will not be a one-size-fits-all endeavor. There are three big themes of interventions have been discussed in the research community: 1) Fair and ethical design and implementation of AI applications, 2) Reducing bias from the algorithm, 3) Ensuring diversity in AI development [15].

3.1 Fair and ethical design and implementation of AI applications

The first theme is fair and ethical design and implementation of AI applications. There are two approaches to this theme, including the establishment of regulatory policies and ethical standards and the promotion of algorithmic transparency. Specifically, the first position is that the policy system for regulated AI development should be strengthened by considering ethical consequences. Researchers argued that a solid policy and educational foundation should be established, including increased awareness, education, stakeholder training, workshops, and the production of educational pamphlets [4]. Next, there was an argument that algorithmic transparency should be guaranteed. This means that decisions made by algorithms should be visible and transparent to the users. The second approach was the most frequently discussed topic when discussing the problem of AI bias [6, 17, 18].

3.2 Reducing bias from the algorithm

The second theme can be divided into detailed approaches: data collection, data preprocessing, development and validation of AI-based algorithms, and model implementation [16]. Nazer et al. [16] emphasized that various types of bias that occur during the data collection process significantly impact the performance and fairness of AI algorithms and that data must be collected that does not distort algorithm results. In particular, the authors highlighted sampling bias resulting from the use of data that was not representative of the entire population and argued that this should be improved. Additionally, they discussed the possibility of bias that can occur in preprocessing, such as data aggregation, missing data handling, and variable selection. They said that bias should not occur in the process of cleaning and processing data after it is collected. Third, they emphasized the importance of appropriate analysis methods and the risk of overfitting during AI algorithm development and validation. The authors explained that overfitting can limit the generalizability of a model, and many AI models often operate in a "black box" manner, hiding biased decision-making processes. The authors argue that thorough validation will be needed to avoid bias and inequality. Lastly, Nazer et al. [16] argued that continuous investment would need to be made to implement AI algorithms in real environments and ensure long-term performance, which may cause model performance to deteriorate over time due to changes in population characteristics and environments. They discussed failure cases due to drift and emphasized the need for continuous monitoring and adjustment to maintain the accuracy and fairness of AI models.

3.3 Ensuring diversity in AI development

The last theme is ensuring diversity in AI development, which includes neutralizing language in data and algorithms

and increasing the number of marginalized groups in AI development [15]. Several scholars have argued that creating cross-disciplinary teams of data scientists and social scientists throughout the process of identifying and addressing bias will be an essential component of the various approaches discussed above [8, 20**?**]. There was also an approach that said data and algorithms should use neutral language to address various issues of bias and discrimination [19]. Lastly, the researchers stated that marginalized groups, including women, people of color, and the LGBTQ+ community, should be actively included in the design and development of AI through user interface, user experience, testing algorithms, and training [22].

4 The Imperative of AI Literacy in Bias Recognition and Avoidance

The paper has looked at how important it is to mitigate and eliminate AI Bias and the efforts being made to do so. Here, while we are putting effort into technological innovation, we will look at why empowering individuals through education is important.

As AI technology is developed and many people use it so easily for their work and studies, various problems arise. They are at risk of losing their critical thinking skills and creativity and relying solely on AI technology. As mentioned earlier, its impact will become even greater without awareness of AI bias. No matter how much technological progress continues, individuals will always be subject to issues related to bias and discrimination inherent in technology. Therefore, individuals must be educated on basic knowledge about why bias occurs in AI, how to recognize it, and how to solve it. It is critical to develop individuals' critical thinking skills and literacy through appropriate training and to empower them to understand the complexities of bias in AI systems.

Equipping people with AI literacy does not mean that AI bias itself can be reduced. However, the impacts and harm of AI bias will be significantly reduced. Empowered individuals will be able to recognize it as bias, evaluate it correctly, and filter it on their own so that it does not interfere with their decision-making even when they encounter it. These empowered individuals will contribute to maintaining a culture that can actively critique and address AI bias. A group of mindful individuals can better recognize instances of bias and advocate for fair and equitable AI systems. Educated AI consumers can make informed choices, question AI-generated results, and demand transparency from technology developers. Efforts to overcome AI bias at the individual level, combined with high-level efforts in the technology sector and academia, will ultimately form a virtuous cycle for AI bias mitigation. Additionally, through engagement with technology developers and policymakers, the community will influence the development and deployment of AI in a more ethical and equitable direction.

Finally, as creators and consumers of AI technology, individuals are ethically responsible for bias and discrimination within AI. They are both creators and consumers of AI technology in the sense that they tailor AI to obtain the information they want through active user interactions. By recognizing and addressing bias in AI, they will be able to align with ethical principles and promote the ethical use of the technology. The landscape of AI and technology is constantly changing, and individuals must continue to learn and adapt to new challenges and biases. Empowered individuals will have the ability to respond to these changes and insist on continuous improvement of AI systems.

5 Essential Considerations and Strategies to Advance AI Literacy

Following Long and Magerko [13]'s definition, AI literacy referred to "a set of competencies that enables individuals to evaluate AI technologies critically; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace." AI literacy encompasses technical proficiency, critical thinking skills, ethical awareness, and a broader understanding of the societal implications of AI technologies. Central to AI literacy is the recognition that AI systems are designed and used by humans, and therefore, individuals must possess the knowledge and skills to navigate AI technologies effectively and responsibly.

I would like to highlight that the groups most in need of AI literacy education are middle and high school students, college students, and young individuals. Based on conversations with several information professionals at the University of Washington Information School, they all emphasized the importance of fostering information literacy among teenagers and university students. They specifically noted that middle and high school students, as well as university students, who spend a significant amount of time on the internet for leisure and assignments, should be the primary targets for AI literacy education. Based on this explanation, students and young individuals are the main stakeholders and target audience for literacy efforts. However, this paper clarifies that the argument applies to all groups who are using or will use AI technology.

The following elements are the most important strategies to prioritize for fostering AI literacy.

5.1 Critical and Ethical Thinking

In order to detect bias in AI, the individual first needs to be unbiased. There will be a need to foster diversity and tolerance by exposing people to case studies from other cultures and regions, as well as discrimination controversies and ethical dilemmas in various fields. In the case of the male doctor and female nurse examples we looked at earlier, if an individual does not see any problems with them at all, he or she will not see any problems with these AI examples.

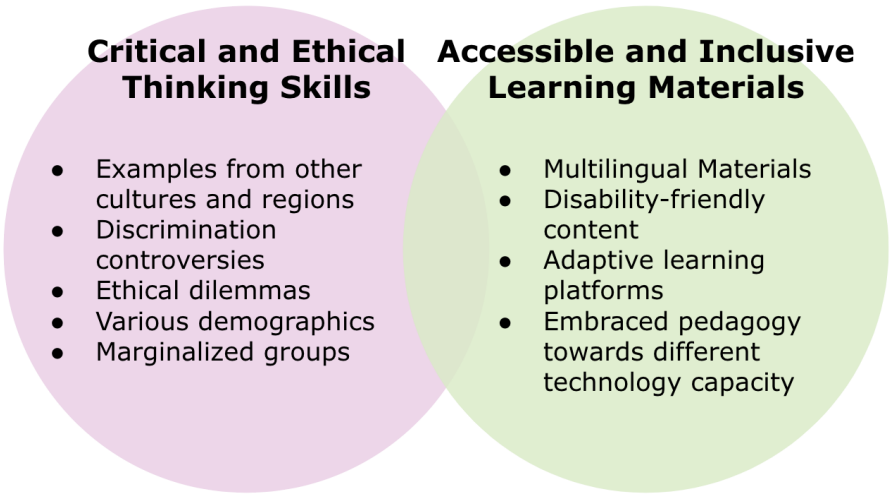

Figure 1. Two Priority Strategies for Fostering AI Literacy

There will be a need to expand the scope of information by exposing students to various viewpoints and many discriminatory debates that have taken place in the past. By presenting learners with a variety of dilemmas and guiding them through the analysis and resolution process, they will be exposed to various perspectives. Additionally, they will have the opportunity to apply diverse critical thinking skills they have learned to real-world AI by assigning tasks that require them to critically analyze AI systems, outputs, and potential biases.

One example implemented in the educational field [9] was to have students test camera detection and AI technology. The students stood in front of the camera and took various poses, such as changing their hairstyles, wearing glasses, and standing in pairs. They after conducted an experiment to evaluate the AI results that analyzed them. In this experiment, which was conducted with students from various racial groups using various experimental poses, uniform results of AI were confirmed, and students were actually able to encounter malfunctions of the algorithm that did not include diversity. This example of an educational experiment is an example in which students directly experienced how AI practices discrimination and does not embrace various representations of real life. The participants expressed worry and anxiety about AI technology and learned how technology can highlight social disparity and injustice.

5.2 Accessible and Inclusive Learning Materials

The strategies discussed above must be comprised of accessible and inclusive learning materials. These should be provided to all through multilingual materials, disability-friendly content, adaptive learning platforms, etc. One of the main problems with bias is the generation of information that does not include some demographic or marginalized groups. If some groups are excluded from education to overcome it, bias and discrimination in various technologies, including AI bias, will continue. AI learning materials should be provided in multiple languages so that non-English speaking users can access them while providing subtitles and transcription services to accommodate learners with different linguistic levels. Disability-friendly content may include screen readers and sign language interpretation services.

In particular, Mohammed and Watson [14] discussed several challenges, saying that equal and accessible AI education should be provided to everyone. They said learners from diverse cultural backgrounds and preferences may not align with the current education system. They discussed that it would be difficult to develop a pedagogy applicable to all of them and provide all students with an equal level of education. In addition, they said that with the recent rapid increase in students' use of AI, mobile devices, and the Internet, it will be difficult to find an educational method that suits each student's digital capacity. However, it was emphasized that efforts should be made to secure inclusive education environments through new perspectives using educational robots and empathic systems.

6 Conclusion

In this position paper, we explore the multifaceted challenges and opportunities for addressing bias and promoting individual empowerment in the context of artificial intelligence (AI). By looking at some examples of what AI bias is, exploring the discourse surrounding its mitigation, understanding the

importance of AI literacy, and what important strategies exist for doing so, we were able to emphasize its important role in educating individuals and shaping the ethical trajectory of AI development. AI bias poses serious ethical and social challenges, requiring cross-disciplinary collaboration and innovative solutions to mitigate it. However, despite various technological efforts, the most important thing is that literacy must be fostered in the information users who consume it. We need to empower them to consume AI information correctly and promote ethical and informed engagement with AI by prioritizing them.

Future research should further deepen the exploration of the Education Framework of AI Literacy by strengthening and supplementing the two strategies claimed in this paper and exploring essential strategies that need to be added. Continued efforts to promote AI literacy through education and support, focusing on fostering critical thinking skills, ethical awareness, and inclusive learning, will require efforts to improve AI literacy at the individual level. As we navigate the rapidly evolving AI technology landscape, stakeholders and individuals must prioritize ethics, their own decision-making process, and human-centric values in AI development and deployment. By fostering a culture of responsible AI engagement based on ethical principles and informed by diverse perspectives, we can ensure that AI technologies contribute to the common good and contribute to a more equitable and socially responsible future.

References

[1] Mirko Bagaric, Jennifer Svilar, Melissa Bull, Dan Hunter, and Nigel Stobbs. 2022. The Solution to the Pervasive Bias and Discrimination in the Criminal Justice System: Transparent and Fair Artificial Intelligence. *American Criminal Law Review* 59 (2022), 95–148. https://heinonline.org/HOL/P?h=hein.journals/amcrimlr59&i=100

[2] Robert Challen, Joshua Denny, Martin Pitt, Luke Gompels, Tom Edwards, and Krasimira Tsaneva-Atansoval. 2019. Artificial intelligence, bias and clinical safety. *BMJ Quality Safety* 28, 3 (Jan. 2019), 231–237. https://doi.org/10.1136/bmjqs-2018-008370

[3] Zhisheng Chen. 2023. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communication* 10 (Sept. 2023), 567. https://doi.org/10.1057/s41599-023-02079-x

[4] Dawson D, Schleiger E, Horton J, McLaughlin J, Robinson C, Quezada G, Scowcroft J, and Hajkowicz S. 2019. Artificial Intelligence: Australia's Ethics Framework. *Data61 CSIRO* (2019), 1–78. https://www.csiro.au/-/media/D61/Reports/Artificial-Intelligence-ethics-framework.pdf

[5] IBM Data and AI Team. 2023. *Shedding light on AI bias with real world examples.* Retrieved May 10, 2024 from https://www.ibm.com/blog/shedding-light-on-ai-bias-with-real-world-examples/

[6] Juan M. Durán and Nico Formanek. 2018. Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines* 28, 4 (Oct. 2018), 645–666. https://doi.org/10.1007/s11023-018-9481-6

[7] Emilio Ferrara. 2024. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Sci* 6, 1 (Dec. 2024), 3. https://doi.org//10.3390/sci6010003

[8] Xavier Ferrer, Tom van Nuenen, Jose M. Such, Mark Coté, and Natalia Criado. 2021. Bias and Discrimination in AI: A Cross-Disciplinary Perspective. *IEEE Technology and Society Magazine* 40, 2 (June 2021), 72–80. https://doi.org/10.1109/MTS.2021.3056293

[9] Georgios Fesakis and Stavroula Prantsoudi. 2021. 3rd European Conference on the Impact of Artificial Intelligence and Robotics. In *Information Processing Management.* 35–42. https://doi.org/10.34190/EAIR.21.039

[10] Sebastian Laacke Charlotte Gauckler. 2023. Bias and Epistemic Injustice in Conversational AI. *THE AMERICAN JOURNAL OF BIOETHICS* 23, 5 (May 2023), 46–48. https://doi.org/10.1080/15265161.2023.2191055

[11] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D. Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. 2023. AI pitfalls and what not to do: mitigating bias in AI. *British Journal of Radiology* 96, 1150 (Oct. 2023). https://doi.org/10.1259/bjr.20230023

[12] Artificial Intelligence (AI) Hub. 2021. *Bias in AI.* Chapman University. Retrieved May 11, 2024 from https://www.chapman.edu/ai/bias-in-ai.aspx

[13] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. *CHI* 20 (April 2020), 598. https://doi.org/10.1145/3313831.3376727

[14] Phaedra S. Mohammed and Eleanor Nell Watson. 2019. Toward Inclusive Education in the Age of Artificial Intelligence: Perspectives, Challenges, and Opportunities. In *Artificial Intelligence and Inclusive Education*, Jeremy Know, Yuchen Wang, and Michael Gallagher (Eds.). 17–37.

[15] Ayesha Nadeem, Babak Abedin, and Olivera Marjanovic. 2020. Gender Bias in AI: A Review of Contributing Factors and Mitigating Strategies. In *ACIS 2020 Proceedings.* AIS Electronic Library (AISeL), 1–13. https://aisel.aisnet.org/acis2020/27?utm_source=aisel.aisnet.org%2Facis2020%2F27&utm_medium=PDF&utm_campaign=PDFCoverPages

[16] Lama H. Nazer, Razan Zatarsh, Shai Waldrip, Janny Xue, Chen Ke, Mira Moukheiber, Ashish K. Khanna, Rachel S. Hicklen, Lama Moukheiber, Dana Moukheiber, Haobo Ma, and Piyush Mathur. 2023. Bias in artificial intelligence algorithms and recommendations for mitigation. *PLOS Digital Health* 2, 6 (2023), e0000278. https://doi.org/10.1371/journal.pdig.0000278

[17] Julian Newman. 2016. Epistemic Opacity, Confirmation Holism and Technical Debt: Computer Stimulation in the Light of Empirical Software Engineering. In *IFIP International Federation for Information Processing 2016.* Springer International Publishing AG, 256–272. https://doi.org/10.1007/978-3-319-47286-7_18

[18] Federica Russo, Eric Schliesser, and Jean Wagemans. 2023. Connecting ethics and epistemology of AI. *AI Society* (Jan. 2023). https://doi.org/10.1007/s00146-022-01617-6

[19] Tianshu Shen, Jiaru Li, Mohamed Reda Bouadjenek, Zheda Mai, and Scott Sanner. 2023. Towards understanding and mitigating unintended biases in language model-driven conversational recommendation. *Information Processing Management* 60, 1 (Jan. 2023), 103139. https://doi.org/10.1016/j.ipm.2022.103139

[20] Helen Sheridan, Emma Murphy, and Dympna O'Sullivan. 2023. Exploring Mental Models for Explainable Artificial Intelligence: Engaging Cross-disciplinary Teams Using a Design Thinking Approach. *Artificial Intelligence in HCI* (July 2023), 337–354. https://doi.org/10.1007/978-3-031-35891-3_21

[21] OpenSpan Team. 2021. *Trustworthy AI: Why We Need It and How to Achieve It.* Retrieved Apr 7, 2024 from https://www.onespan.com/blog/trustworthy-ai-why-we-need-it-and-how-achieve-it

[22] Meredith Whittaker, Meryl Alper, Cynthia L. Bennett, Sata Hendren, Liz Kaziunas, Mara Mills, Meredith Ringel Morris, Joy Rankin, Emily Rogers, Marcel Salas, and Sarath Myers West. 2019. Disability, Bias, and AI.