# Promoting Equity in AI-Driven Mental Health Care for Marginalized Populations

**Nii Tawiah[1], Judith P. Monestime[2]**
[1]Delaware State University
[2]Florida Atlantic University
ntawiah@desu.edu, jmonestime@fau.edu

**Abstract**

Artificial Intelligence (AI) is increasingly used in mental health care, but its equitability is a pressing concern. This paper examines the potential biases in AI-driven mental health tools and their impact on marginalized communities. It explores several strategies to mitigate bias in AI-driven mental tools, focusing on promoting equity and inclusivity.

## Introduction

The increasing integration of Artificial Intelligence (AI) in mental health care has introduced innovative avenues for diagnosing, treating, and monitoring mental health conditions. Despite its potential, the proliferation of AI-driven mental health tools has prompted concerns regarding bias.

This concern is particularly significant as biased AI algorithms can cause unequal treatment and outcomes, posing a considerable threat to underserved groups. This paper delves into the origins of bias in AI mental health tools and its implications for patient care. It proposes mitigation strategies to guarantee the equitable benefit of these technologies for all individuals.

## Bias in AI-driven Mental Health Tools

AI algorithms are only as good as the data they are trained on, and if this data is biased, the outcomes will be too (Bolukbasi et al. 2016). For instance, a study found that facial recognition technology, used in some mental health diagnosis tools, had a higher error rate for darker-skinned individuals (Raji 2020).

## Sources of Bias in AI Mental Health Tools

Bias in AI-driven mental health tools can stem from several sources. The most common source is the data used to train AI models. If the training data does not represent the diverse populations the tools are intended to serve, the AI models may not perform equally well across different demographic groups. For instance, if a mental health tool is trained predominantly on data from white, middle-class individuals, it may not accurately diagnose or treat conditions in individuals from other racial demographics (Buolamwini and Gebru 2018).

Another source of bias is the design and development process of AI tools. If the development team lacks diversity or fails to consider the specific needs of various populations, the resulting tools may inadvertently reflect the biases of their creators. This can lead to tools that are less effective or even harmful for certain groups, exacerbating existing disparities in mental health care (Gebru et al. 2020).

Moreover, the algorithms themselves can introduce bias. For example, machine learning algorithms often optimize for accuracy or efficiency without considering fairness. As a result, they may reinforce existing biases in the data or even create new biases through their decision-making processes (Hardt et al. 2016).

## Impact of Bias on Patient Care

The presence of bias in AI-driven mental health tools can have significant consequences for patient care. Biased AI tools may misdiagnose conditions in certain populations, leading to inappropriate or ineffective treatments. For example, a tool that underdiagnoses depression in Black patients compared to white patients could contribute to ongoing disparities in mental health outcomes (Obermeyer et al. 2019).

## Mitigating Bias in AI Mental Health Tools

Addressing bias in AI-driven mental health tools is crucial to ensuring that these technologies provide equitable care. One approach is to use diverse and representative datasets during the training phase. By including data from a wide range of populations, developers can create AI models that

perform better across different demographic groups (Chen et al. 2020).

Additionally, involving diverse stakeholders in the design and development process can help identify potential biases and ensure that the tools meet the needs of all users. This includes technical experts, mental health professionals, patients, and advocates from various backgrounds (Vollmer et al. 2020).

Furthermore, developers can implement fairness-aware algorithms that aim to reduce AI decision-making bias. Techniques such as reweighting, fairness constraints, and algorithmic audits can help ensure that AI tools make equitable decisions (Bellamy et al. 2019).

## Impact on Marginalized Communities

AI applications in mental health have demonstrated high accuracies in predicting, classifying, and subgrouping mental illnesses using various data sources (Graham et al. 2019). However, language-based AI models exhibit significant biases regarding religion, race, gender, nationality, sexuality, and age (Straw and Callison-Burch 2020). These biases could reinforce and perpetuate existing inequities in mental health care if not addressed.

## Disparities in AI-Driven Mental Health Care

Vulnerable populations already face significant disparities in mental health care, including reduced access to services, stigma, and culturally insensitive care. The introduction of AI-driven tools can either exacerbate these disparities or help mitigate them, depending on how these technologies are developed and implemented.

One of the critical issues is the risk of bias in AI algorithms. AI systems are trained on large datasets, and if these datasets are not representative of diverse populations, the resulting models may not perform well for all groups. For instance, AI tools trained primarily on data from white, middle-class individuals may not accurately diagnose or treat mental health conditions in people from other racial or socioeconomic backgrounds (Obermeyer et al. 2019). This can lead to misdiagnoses, inappropriate treatments, and overall poorer mental health outcomes for underserved groups.

## Targeted Effects on Minority Groups

***Racial and Ethnic Minorities:*** AI-driven mental health tools can perpetuate existing racial biases in healthcare. For example, suppose an AI tool is trained on data that underrepresents Black or Hispanic individuals. In that case, it may be less accurate in diagnosing conditions like depression or anxiety in these groups. This can result in delayed or incorrect diagnoses, further entrenching disparities in mental health outcomes (Buolamwini and Gebru, 2018).

***Low-Income Populations:*** AI tools that require high levels of digital literacy or access to advanced technology may not be accessible to low-income individuals, who are less likely to have access to such resources. This digital divide can prevent these populations from benefiting from AI-driven mental health tools, widening the gap in mental health care (Eubanks 2018).

***Individuals with Disabilities:*** AI tools not designed with accessibility in mind may fail to serve individuals with disabilities adequately. For example, tools that rely heavily on written input or visual cues may be ineffective for individuals with cognitive or sensory impairments. This can lead to the exclusion of people with disabilities from the benefits of AI in mental health care (Shinohara and Wobbrock 2011).

***Non-English-Speaking Immigrants*:** Non-English-speaking immigrants face significant barriers in accessing mental health services, a challenge that extends to AI-driven interventions. Most AI models in mental health are trained on English-language datasets, which may not apply to non-English-speaking populations (Zhou et al. 2021). The lack of culturally and linguistically appropriate AI tools can result in erroneous diagnoses or ineffective treatment recommendations. To address this issue, AI systems must be developed with multilingual capabilities and cultural sensitivity to ensure accurate and equitable mental health care for non-English speaking immigrants.

***Adolescents*:** Adolescents represent a vulnerable population in mental health care, with unique needs and challenges that differ from those of adults. AI-driven mental health tools designed for general populations may not be effective for adolescents due to differences in communication styles, cognitive development, and mental health risk factors (Luxton et al. 2011). Developing age-specific AI models that account for these differences is crucial, as well as providing tailored interventions that resonate with adolescents. Additionally, involving adolescents in the design and testing of AI tools can help create more effective and user-friendly solutions.

***Residents of Rural Communities*:** Residents of rural communities often face a lack of access to mental health services, a problem that AI has the potential to mitigate. However, the digital divide—referring to disparities in internet access and technological infrastructure—poses a significant barrier to the effective implementation of AI-driven mental health care in these areas (Torous et al. 2021). To promote equity, strategies must be developed to ensure that rural residents have access to technology and digital literacy support to benefit from AI interventions. This includes investment in broadband infrastructure and providing low-cost or free digital devices to underserved populations.

## Mitigating Negative Impacts

To mitigate the negative impacts of AI-driven mental health tools on socially excluded communities, several strategies can be employed:

***Inclusive Data Practices:*** Ensuring that AI models are trained on diverse and representative datasets is crucial. This involves actively seeking out data from underrepresented groups and being mindful of potential biases in data collection and labeling processes (Gebru et al. 2020).

***Stakeholder Involvement:*** Engaging with underrepresented communities while developing and implementing AI tools can help ensure that these technologies meet their needs. This includes involving community leaders, mental health professionals, and individuals from these communities in the design process (Vollmer et al. 2020).

***Accessibility and Fairness:*** Developing AI tools that are accessible to individuals with different levels of digital literacy and those with disabilities is essential. Fairness-aware algorithms can minimize bias in decision-making processes and ensure equitable outcomes across various groups (Hardt et al. 2016).

## Solutions

To address these concerns, transparency, and accountability in the AI decision-making must be emphasized (Bolukbasi et al. 2016). There must be a need for coordinated training and education programs for mental health professionals to improve trust in AI solutions (Viswanathan et al. 2022; Zhang et al. 2023).

***Diverse and Representative Datasets:*** One of the most effective ways to reduce bias in AI-driven mental health tools is to ensure that the training data is diverse and representative of the populations the tools are intended to serve. Biases often arise when AI models are trained on data that does not adequately reflect the diversity of the population, leading to models that perform poorly for underrepresented groups.

To create more inclusive datasets, data collection efforts must include individuals from various racial, ethnic, socioeconomic, and cultural backgrounds. This includes collaborating with community organizations, healthcare providers, and patients to gather data that accurately represent marginalized communities' diverse experiences and needs (Gebru et al. 2020). Additionally, it is essential to continually update datasets to reflect changing demographics and ensure that AI models remain relevant and practical.

***Fairness-Aware Algorithms:*** Implementing fairness-aware algorithms is another critical step in addressing bias in AI-driven mental health tools. These algorithms are designed to identify and mitigate bias during the model training process, ensuring that the AI system makes fair and equitable decisions across different groups.

One approach to fairness-aware algorithms uses techniques such as reweighting, which adjusts the importance of other data points to ensure that underrepresented groups are given equal consideration in the model's predictions (Zemel et al. 2013).

Another technique is using fairness constraints, which explicitly incorporate fairness objectives into the model's optimization process (Hardt et al. 2016). These methods help to reduce disparities in outcomes and ensure that AI tools provide equitable care for all individuals.

***Inclusive Design and Development:*** Involving diverse stakeholders in the design and development process of AI-driven mental health tools is crucial to ensuring that these technologies are inclusive and meet the needs of marginalized communities. This includes engaging with patients, mental health professionals, community leaders, and advocates from various backgrounds to gather input on the design, functionality, and potential impact of the tools.

By incorporating diverse perspectives, developers can identify potential biases early in the development process and design AI tools that are culturally sensitive, accessible, and responsive to the needs of different populations (Vollmer et al. 2020).

***Algorithmic Transparency and Accountability:*** Algorithmic transparency involves making the decision-making processes of AI systems understandable and accessible to users, developers, and regulators. This includes providing clear explanations of how AI models make predictions and allowing stakeholders to scrutinize the models for potential biases. Implementing mechanisms for algorithmic accountability is also crucial. This can involve regular audits of AI systems to detect and correct biases, as well as establishing guidelines and regulations to ensure that AI tools are used ethically and fairly (Mitchell et al. 2019).

By fostering transparency and accountability, developers and organizations can build trust in AI-driven mental health tools and ensure that they are used to promote equity in mental health care.

***Education and Training for Developers and Practitioners:*** Education and training programs for developers should emphasize the importance of diversity, equity, and inclusion in AI development and provide practical strategies for creating unbiased AI systems (Holstein et al. 2019).

Similarly, mental health practitioners who use AI-driven tools must be trained to understand the limitations of these technologies and to recognize potential biases in their outputs. Repositioning AI as a decision support tool rather than an absolute decision-maker may lead to improved acceptance and adoption (Viswanathan et al. 2022; Grzenda 2021).

This includes being aware of how biases in AI tools might affect diagnosis and treatment decisions and taking steps to ensure that AI complements rather than replaces human judgment.

## Conclusion

The potential of AI-driven mental health tools to enhance mental health care is substantial, but their impact on disenfranchised communities necessitates thorough consideration. In the absence of deliberate efforts to address bias, accessibility, and inclusivity, these tools run the risk of perpetuating existing disparities in mental health care. Through the adoption of inclusive practices and continued engagement with disadvantaged groups during the AI development process, there exists the potential to forge tools that advance fair mental health outcomes for all.

The presence of bias in AI-driven mental health tools stands as a formidable impediment to the realization of equitable mental health care. By acknowledging and rectifying the roots of bias, developers, and practitioners can strive toward crafting AI tools that effectively cater to all demographics. Sustained efforts in this sphere are necessary to ensure that the advantages of AI in mental health care are within reach of all, irrespective of their backgrounds.

Confronting bias in AI-driven mental health tools is critical to guaranteeing the equitable benefit of these technologies to all individuals. Through the utilization of diverse and representative datasets, the implementation of fairness-aware algorithms, the active involvement of a diverse array of stakeholders in the design process, the promotion of transparency and accountability, and the provision of education and training, it is plausible to mitigate bias and fabricate AI tools that champion equity in mental health care. These remedies are pivotal in fully harnessing the potential of AI to facilitate mental health outcomes, particularly for marginalized communities.

## Acknowledgments

The authors would like to thank the South Big Data Innovation Hub for their support through the South Hub Partnership Nucleation Award.

## References

Bellamy, R. K., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., ... & Zhang, Y. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, *63*(4/5), 4-1.

Bolukbasi, T., Chang, K. W., Zou, J. Y., Saligrama, V., & Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, *29*.

Buolamwini, J., & Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency* (pp. 77-91). PMLR.

Chen, I. Y., Szolovits, P., & Ghassemi, M. 2019. Can AI help reduce disparities in general medical and mental health care? *AMA journal of ethics*, *21*(2), 167-179.

Eubanks, V. 2018. *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.

Grote, T., & Berens, P. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of medical ethics*, *46*(3), 205-211.

Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H. C., & Jeste, D. V. 2019. Artificial Intelligence for Mental Health and Mental Illnesses: an Overview. *Current psychiatry reports*, *21*(11),116.

Grzenda, A. 2021. Artificial intelligence in mental health. *Converg. Ment. Heal. A Transdiscipl. Approach to Innov*, 137.

Hardt, M., Price, E., & Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.

Holstein, K., Wortman Vaughan, J., Daumé III, H., Dudik, M., & Wallach, H. 2019. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1-16).

Luxton, D. D., June, J. D., & Kinn, J. T. 2011. Technology-based suicide prevention: current applications and future directions. *Telemedicine and e-Health*, *17*(1), 50-54.

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency* (pp. 220-229).

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

Raji, I. D. 2020. Saving Face: Investigating the Impact of Racial and Gender Bias in Facial Recognition Technology. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency*, 331-341.

Shinohara, K., & Wobbrock, J. O. 2011. In the shadow of misperception: assistive technology use and social interactions. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 705-714).

Straw, I., & Callison-Burch, C. 2020. Artificial Intelligence in mental health and the biases of language based models. *PLOS ONE*, *15*(12), e0240376.

Torous, J., Bucci, S., Bell, I. H., Kessing, L. V., Faurholt-Jepsen, M., Whelan, P., ... & Firth, J. 2021. The growing field of digital psychiatry: current evidence and the future of apps, social media, chatbots, and virtual reality. *World Psychiatry*, *20*(3), 318-335.

Viswanathan, S., Rollwage, M., & Harper, R. 2022. Promises and Challenges of AI-Enabled Mental Healthcare: A Foundational Study. In *Empowering Communities: A participatory approach to AI for mental health*.

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ... & Hemingway, H. 2020. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, *368*.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. 2013. Learning fair representations. In *International conference on machine learning* (pp. 325-333). PMLR.

Zhang, M., Scandiffio, J., Younus, S., Jeyakumar, T., Karsan, I., Charow, R., ... & Wiljer, D. 2023. The Adoption of AI in Mental Health Care–Perspectives From Mental Health Professionals: Qualitative Descriptive Study. *JMIR Formative Research*, *7*(1), e47847.

Zhou, X., Snoswell, C. L., Harding, L. E., Bambling, M., Edirippulige, S., Bai, X., & Smith, A. C. 2020. The role of telehealth in reducing the mental health burden from COVID-19. *Telemedicine and e-Health*, *26*(4), 377-379.