International Journal of Physical and Social Sciences
Vol.15 Issue 1, Jan 2025,
ISSN: 2249-5894 Impact Factor: 6.644
Journal Homepage: http://www.ijmra.us, Email: editorijmie@gmail.com
Double-Blind Peer Reviewed Refereed Open Access International Journal - Included in the International Serial Directories Indexed & Listed at: Ulrich's Periodicals Directory ©, U.S.A., Open J-Gate as well as in Cabell’s Directories of Publishing Opportunities, U.S.A

AI Bias and Its Psychological Impact on Marginalized Groups: A Critical Examination of Algorithmic Injustice

Dr. Ruma kumari Sinha
Assistant Professor,
Department of Psychology
S.K.M. College, Begusarai, L. N. M. U. Darbhanga, Bihar

Abstract

Artificial intelligence (AI) systems are increasingly used in high-stakes decision-making across domains such as policing, healthcare, hiring, and education. However, these systems often reproduce and amplify existing social biases due to skewed training data, flawed model assumptions, or lack of diversity in design. While the technical dimensions of AI bias have received considerable attention, its psychological impact on marginalized groups remains critically underexplored.

This paper examines how algorithmic bias affects the mental health and psychological well-being of individuals from marginalized communities, including racial minorities, women, LGBTQ+ individuals, and people with disabilities. Drawing on psychological theories such as stereotype threat, minority stress, and learned helplessness, we explore how biased AI systems contribute to increased anxiety, reduced self-esteem, identity conflict, and institutional mistrust. Real-world case studies—including wrongful arrests caused by facial recognition errors and discriminatory hiring algorithms—demonstrate the harmful outcomes of algorithmic injustice.

We argue that these technologies, when biased, act not only as tools of discrimination but also as psychological stressors that reinforce feelings of exclusion and powerlessness. The paper calls for an expanded ethical framework in AI development that considers psychological harm as a measurable consequence of bias. Mitigation strategies such as inclusive data practices, participatory design, and psychological harm assessments are discussed.

Ultimately, this research highlights the urgent need to integrate psychological insights into AI ethics, governance, and design. Addressing the mental and emotional toll of biased AI is essential to building systems that are truly fair, transparent, and socially responsible.

Keywords: AI Bias**,** Algorithmic Discrimination**,** Marginalized Communities**,** Psychological Impact**,** Mental Health**,** Trust in Technology**,** Algorithmic Injustice**,** Ethical AI

Introduction

In recent years, artificial intelligence (AI) has transitioned from a futuristic concept to a pervasive force shaping many aspects of daily life. From predictive policing and automated hiring tools to loan approvals and medical diagnostics, AI systems are increasingly used to inform and automate critical decisions that directly impact human lives. While these technologies promise efficiency, consistency, and scalability, they also carry a significant and often overlooked risk: **algorithmic bias**—the systematic production of unfair or discriminatory outcomes by AI systems.

Algorithmic bias arises from a variety of sources, including biased training data, flawed model assumptions, and design processes that exclude the perspectives of marginalized populations. Despite being framed as objective or neutral, AI systems can reflect and reinforce historical inequalities embedded within society. Facial recognition systems misidentifying Black individuals, hiring algorithms that disadvantage women, and healthcare models that under-prioritize the needs of Black patients are not isolated incidents—they are symptoms of deeper structural problems in how AI is developed and deployed.

Most existing research on AI bias has focused on technical solutions—how to detect, quantify, and mitigate bias within algorithms. While this work is vital, it leaves a critical question unanswered: **What happens to the people affected by these biases?** Specifically, what are the **psychological consequences** for individuals and communities who are repeatedly disadvantaged by so-called intelligent systems?

Marginalized groups—such as racial and ethnic minorities, women, LGBTQ+ individuals, people with disabilities, and low-income populations—are disproportionately impacted by biased AI systems. These groups already experience systemic inequalities and discrimination in social, political, and economic contexts. When AI technologies amplify these injustices, the effects are not only material but also psychological. Repeated exposure to algorithmic discrimination can contribute to **mental health issues**, including increased stress, anxiety, depression, and identity-related conflict. Moreover, it can erode **trust in institutions**, discourage civic participation, and reinforce a sense of social exclusion or alienation.

Understanding the **psychological dimensions of AI bias** requires moving beyond technical evaluations of accuracy or fairness. Instead, it demands an interdisciplinary lens that incorporates insights from psychology, sociology, and ethics. Psychological theories such as **stereotype threat**, **learned helplessness**, and the **minority stress model** provide valuable frameworks for understanding how biased outcomes affect individual well-being, self-perception, and long-term mental health. These frameworks help explain how seemingly small algorithmic decisions can lead to large-scale emotional harm—especially when individuals feel powerless to contest or even understand the decisions made about them.

Consider the experience of a job seeker who is repeatedly rejected by automated hiring systems that devalue resumes from women or ethnic minorities. Over time, these rejections can foster feelings of inadequacy, hopelessness, and self-doubt—even if the individual is fully qualified. Similarly, a young Black man wrongfully identified by facial recognition software as a criminal suspect may not only face legal consequences but also long-term trauma, fear of surveillance, and mistrust in the justice system. These are not just failures of technology—they are deeply **human harms** with lasting psychological implications.

In many cases, the **invisibility and opacity** of AI systems exacerbate the psychological impact. Unlike interpersonal discrimination, which can be identified and sometimes challenged, algorithmic decisions are often hidden behind proprietary models and technical jargon. This lack of transparency makes it difficult for affected individuals to understand why a decision was made or to seek redress. The resulting **loss of agency** can increase feelings of helplessness and distress, particularly among populations already marginalized by other societal forces.

Despite these significant concerns, psychological impacts are rarely considered in mainstream discussions of AI ethics or policy. Current frameworks tend to emphasize fairness, accountability, and transparency—but often neglect emotional well-being, identity validation, and the mental health effects of exclusion. This omission represents a critical gap in the growing field of AI ethics and governance. As AI continues to evolve and permeate everyday life, addressing this gap becomes increasingly urgent.

This paper aims to fill that gap by critically examining the psychological effects of AI bias on marginalized communities. Drawing on empirical research, theoretical models, and real-world case studies, we explore how biased AI systems contribute to emotional distress, identity harm, and institutional mistrust. We argue that **algorithmic bias must be understood not only as a technical flaw or ethical issue—but as a form of psychological harm** that reinforces social exclusion and undermines individual well-being.

In doing so, we propose a more holistic and human-centered approach to AI governance—one that considers the lived experiences and emotional realities of those most impacted by these technologies. We also discuss practical strategies for mitigating these harms, including participatory design, inclusive data practices, and the integration of psychological harm assessments into algorithmic audits. By incorporating psychological insights into the evaluation and regulation of AI systems, we can begin to develop technologies that are not only intelligent but also just, inclusive, and emotionally safe for all users.

In sum, this research seeks to shift the conversation from "How do we make AI fair?" to "How do we ensure AI does not harm people psychologically—especially those who are already at risk?" It is only through this expanded lens that we can truly create ethical AI systems that serve society equitably and compassionately.

Methodology

This study adopts a **qualitative, interdisciplinary research approach** to examine the psychological effects of AI bias on marginalized groups. Given the complex and multifaceted nature of the topic—intersecting technology, psychology, and ethics—a mixed-methods empirical study was considered but ultimately excluded due to limitations in available primary data. Instead, this paper employs a **critical literature review and thematic content analysis** of existing case studies, peer-reviewed psychological research, and documented instances of algorithmic bias in real-world applications.

4.1 Research Design

The research design is structured in three phases:

1. **Literature Review**

   A comprehensive review of scholarly literature was conducted across databases such as **PsycINFO**, **PubMed**, **Google Scholar**, and **IEEE Xplore**, focusing on three key domains:

- Psychological theories and effects of discrimination (e.g., stereotype threat, minority stress)
- Empirical studies on algorithmic bias in AI systems
- Ethical and policy-focused papers on AI governance and fairness

2. **Case Study Selection**

   Real-world incidents of AI bias were selected based on public accessibility, relevance to marginalized populations, and available documentation. These include:

   - Facial recognition misidentification of Black individuals in law enforcement
   - Discriminatory hiring algorithms disadvantaging women and ethnic minorities
   - Healthcare algorithms that under-prioritize treatment for Black patients

3. **Thematic Content Analysis**

   A thematic analysis approach was used to extract recurring psychological and emotional themes from the literature and case studies. Themes were organized into categories such as:

   - Mental health outcomes (e.g., anxiety, depression, stress)
   - Identity-related impacts (e.g., self-worth, stereotype internalization)
   - Behavioral responses (e.g., withdrawal from systems, institutional mistrust)

4.2 Inclusion Criteria

- Studies and sources published between **2015 and 2024**
- Peer-reviewed journal articles, technical reports, and ethical white papers
- Case studies involving **marginalized groups** (e.g., racial minorities, women, LGBTQ+ individuals, people with disabilities)
- Sources with a clear link between algorithmic bias and **individual or collective psychological outcomes**

Inclusion Criteria

For this research on **AI bias and its psychological impact on marginalized groups**, sources and data were included based on the following criteria:

- **Publication Date:** Studies and sources published between **January 2015 and December 2024** to capture the most recent and relevant developments in AI bias and psychological research.
- **Relevance:** Research explicitly addressing the intersection of **algorithmic bias** or AI-related discrimination and **psychological outcomes** such as mental health, identity, trust, or emotional well-being.
- **Population Focus:** Studies involving or focusing on **marginalized or underrepresented groups**, including but not limited to racial and ethnic minorities, women, LGBTQ+ individuals, people with disabilities, and low-income populations.
- **Research Type:** Peer-reviewed journal articles, conference proceedings, technical reports, systematic reviews, meta-analyses, and well-documented case studies.
- **Language:** Publications available in **English** to ensure accurate interpretation and analysis.
- **Data Transparency:** Sources providing sufficient methodological detail or data transparency to support analysis, including descriptions of AI systems, bias identification methods, and psychological impact assessments.

4.3 Exclusion Criteria

- Technical papers without a psychological or sociological component
- Studies focused solely on computational methods of bias detection
- Sources with insufficient methodological transparency or bias documentation

4.4 Theoretical Framework

This study is grounded in established **psychological and sociological theories**, including:

- **Stereotype Threat Theory** (Steele & Aronson, 1995)
- **Minority Stress Theory** (Meyer, 2003)
- **Learned Helplessness Theory** (Seligman, 1975)
- **System Justification Theory** (Jost & Banaji, 1994)

These frameworks guide the interpretation of AI bias not simply as technical error but as a **psychosocial process** with potential long-term effects on individual identity, emotional regulation, and social engagement.

4.5 Ethical Considerations

As this study relies solely on secondary data and publicly available case documentation, it does not involve direct interaction with human subjects and does not require institutional review board (IRB) approval. However, **ethical sensitivity** is maintained throughout by avoiding victim-blaming, ensuring respectful language, and highlighting the voices of those affected by algorithmic injustice.

4.6 Limitations

- The absence of primary data collection limits the ability to generalize findings across all affected populations.
- Thematic analysis is inherently interpretive and may be influenced by researcher bias, though triangulation with existing theories was used to enhance validity.
- The fast-evolving nature of AI may mean newer technologies or cases are not yet documented or available in academic literature.

6. Results and Discussion

6.1 Results

The data reviewed from 2015 to 2024 reveal persistent and systemic **AI bias** impacting marginalized groups across multiple sectors, including healthcare, criminal justice, employment, and digital communication platforms. The empirical evidence shows that AI systems frequently inherit and amplify existing social inequalities due to biased training data, lack of diverse representation in design teams, and insufficient regulatory oversight.

Key findings include:

- **Facial Recognition and Racial Bias:** Multiple studies documented significantly higher error rates for facial recognition algorithms when identifying Black and darker-skinned individuals compared to white counterparts. This disparity has led to wrongful arrests and increased surveillance of marginalized communities, causing psychological distress manifesting as anxiety, mistrust in law enforcement, and feelings of social alienation.
- **Employment and Gender Bias:** AI-powered hiring tools have been shown to discriminate against women and ethnic minorities by prioritizing historical hiring data biased toward majority groups. This exclusion contributes to diminished self-esteem, workplace marginalization, and increased stress among affected candidates.

- **Healthcare Disparities:** Healthcare AI algorithms often underdiagnose or undertreat marginalized groups due to limited diversity in clinical datasets. For example, pulse oximeters inaccurately measure oxygen saturation in individuals with darker skin, and chatbots perpetuate medically racist stereotypes. These biases contribute to poorer health outcomes, increased medical mistrust, and psychological stress among patients.
- **Mental Health AI Tools:** Emerging research highlights racial disparities in AI models used for mental health diagnostics. Tools designed to detect depression from social media posts showed significantly lower accuracy for Black users, risking underdiagnosis and lack of access to care.
- **Public Attitudes Toward AI:** Survey data reveal that marginalized populations, including gender minorities and neurodivergent individuals, express higher skepticism and fear toward AI systems. This distrust is linked to lived experiences of bias, discrimination, and exclusion, which negatively impact psychological well-being and willingness to engage with AI technologies.

6.2 Discussion

These results underscore the urgent need to view AI bias not merely as a technical problem but as a **complex psychosocial phenomenon** with far-reaching consequences for marginalized groups' mental health and social integration.

Psychological Impact and Identity

The evidence indicates that AI bias functions as a modern form of systemic discrimination, reinforcing societal stereotypes and power imbalances. The mental health consequences—such as anxiety, depression, and identity threat—align with psychological theories including **stereotype threat** and **minority stress theory**. For instance, repeated misidentification by facial recognition systems may induce a sense of vulnerability and loss of agency, while biased hiring algorithms contribute to feelings of exclusion and diminished self-worth.

The internalization of bias—where individuals begin to believe and accept negative stereotypes about their group—can exacerbate these outcomes, further harming emotional well-being. Such internalized stigma may also reduce marginalized individuals' motivation to participate in institutions that rely heavily on AI, including healthcare, employment, and law enforcement.

Trust and Institutional Skepticism

AI bias erodes trust in social and technological institutions. When AI systems disproportionately harm marginalized groups, it fosters **institutional mistrust**, a critical barrier to equitable access to services. For example, health disparities caused by biased algorithms may deter patients from seeking care or adhering to medical advice, worsening health inequities. Similarly, mistrust in AI-driven law enforcement tools can decrease community cooperation, negatively impacting public safety.

Addressing this trust deficit requires transparent AI design, inclusive stakeholder engagement, and policies that prioritize fairness and accountability.

Intersectionality and Layered Vulnerabilities

The results highlight how AI bias intersects with multiple dimensions of identity—race, gender, disability, and socioeconomic status—creating layered vulnerabilities. Individuals who belong to multiple marginalized groups face compounded effects of bias, making them particularly susceptible to psychological harm.

This intersectional lens is crucial for developing AI systems and interventions that are sensitive to diverse lived experiences rather than applying one-size-fits-all solutions.

Moving Toward Ethical and Anti-Racist AI

The data reveal a growing movement toward integrating **anti-racist principles** and psychological considerations into AI governance. Recent proposals advocate for incorporating frameworks such as **racism-related stress** models into psychiatric AI tools to better serve marginalized communities. These approaches emphasize the need to embed fairness, transparency, and cultural competence within AI systems from development through deployment.

Moreover, regulatory efforts must mandate routine bias audits, participatory design with affected communities, and mechanisms for redress when harms occur.

6.3 Limitations

While the reviewed studies offer valuable insights, several limitations constrain the generalizability of findings. Most data rely on secondary sources and case studies from Western contexts, primarily the U.S. and Europe, limiting understanding of AI bias impacts globally. Additionally, few studies have longitudinal designs to assess the long-term psychological effects of AI bias.

The interpretive nature of thematic analysis introduces potential researcher bias, though triangulation with established psychological theories mitigates this risk. Finally, the rapidly evolving AI landscape means newer biases may emerge that are not yet documented.

6.4 Implications and Future Research

The findings call for multidisciplinary research combining computer science, psychology, sociology, and ethics to holistically address AI bias. Future work should prioritize:

- **Longitudinal studies** assessing mental health trajectories in affected populations.
- **Participatory research** involving marginalized communities in AI design and evaluation.
- **Development of bias mitigation tools** that explicitly incorporate psychological impact assessments.
- **Policy frameworks** that enforce transparency, accountability, and equitable AI deployment.

Conclusion

Artificial Intelligence (AI) has emerged as a transformative force across nearly every sector of society, promising innovations that range from automating mundane tasks to revolutionizing healthcare, criminal justice, employment, and communication. However, as this study has demonstrated, the deployment of AI systems has also unveiled profound challenges—most notably, the persistent presence of **AI bias** and its consequential **psychological impact on marginalized groups**. Between 2015 and 2024, the accumulation of evidence from diverse domains consistently points to a troubling pattern: AI technologies, far from being neutral or objective, often replicate and intensify historical social inequalities, thereby imposing psychological burdens on already vulnerable populations.

AI Bias as a Socio-Technical Problem

At the heart of this issue lies a fundamental misunderstanding of AI as a purely technical innovation. This research highlights that AI bias is not solely a matter of flawed algorithms or incomplete datasets but a complex **socio-technical phenomenon** embedded within broader societal structures and power dynamics. Marginalized groups—including racial and ethnic minorities, women, LGBTQ+ individuals, people with disabilities, and low-income populations—are disproportionately targeted by AI bias due to the ways these technologies are designed, trained, and deployed within contexts steeped in systemic discrimination.

The technical origins of bias are multifaceted: biased training data that overrepresent dominant groups, lack of diversity among AI developers, insufficient regulatory standards, and limited accountability mechanisms. Yet, this study emphasizes that focusing only on technical fixes obscures the **human costs** these biases exact—manifesting in measurable psychological harms such as anxiety, depression, identity threat, diminished self-esteem, and pervasive mistrust.

Psychological Impact and Theoretical Implications

The psychological consequences of AI bias are profound and resonate with established theories in social psychology and mental health. The phenomenon of **stereotype threat**, where individuals underperform or experience distress due to awareness of negative stereotypes about their group, provides a crucial lens through which to understand AI bias's emotional toll. For instance, repeated misidentification by facial recognition systems or rejection by biased hiring algorithms may trigger heightened stress responses, feelings of vulnerability, and diminished self-efficacy.

Similarly, the **minority stress model**—which explains how chronic exposure to social stigma leads to adverse health outcomes—can be applied to AI contexts. Marginalized individuals facing AI discrimination endure cumulative stressors that compound existing inequalities, exacerbating mental health disparities. Moreover, the internalization of negative stereotypes, or **internalized stigma**, further undermines psychological well-being, fostering a sense of helplessness and social exclusion.

These theoretical frameworks underscore the need to view AI bias through a psychosocial lens that acknowledges how technology-mediated discrimination intertwines with identity and mental health.

Domains of Impact: Evidence Across Sectors

This study synthesized data across several critical sectors where AI bias has direct psychological repercussions:

- **Criminal Justice and Surveillance:** Facial recognition technologies with racial bias have resulted in wrongful arrests and heightened surveillance of Black and minority communities. These incidents not only cause immediate legal and social harms but also induce chronic anxiety, fear, and erosion of trust toward law enforcement, disrupting community relations and individual mental health.
- **Employment:** AI-based recruitment tools trained on biased historical data tend to favor majority groups, systematically disadvantaging women and ethnic minorities. Such exclusion contributes to psychological distress, decreased job satisfaction, and feelings of alienation, further entrenching socioeconomic disparities.
- **Healthcare:** AI algorithms in medical diagnostics and treatment often lack sufficient representation of marginalized populations, leading to underdiagnosis, misdiagnosis, and inequitable care. These biases erode patient trust, provoke anxiety regarding medical treatment, and exacerbate health inequalities, with significant psychological fallout.
- **Mental Health Technologies:** AI-driven tools designed to detect mental health conditions, such as depression, often perform poorly for marginalized groups due to cultural and linguistic biases in training data. This results in underdiagnosis and

inadequate treatment, compounding mental health challenges in already vulnerable populations.

- **Public Perception and Trust:** Surveys reveal that marginalized groups express greater skepticism and fear toward AI technologies, stemming from lived experiences of discrimination. This mistrust undermines the potential benefits of AI and creates barriers to equitable technology adoption.

Intersectionality and Compounded Vulnerabilities

A key insight from this research is the necessity of adopting an **intersectional framework** when addressing AI bias. Individuals often embody multiple marginalized identities, and the effects of AI bias are not additive but **multiplicative**, leading to compounded vulnerabilities. For example, a Black woman with a disability may face overlapping biases from facial recognition, employment algorithms, and healthcare AI simultaneously, magnifying psychological harm.

Ignoring intersectionality risks developing narrow solutions that fail to address the nuanced realities of marginalized individuals. Instead, AI systems and policy interventions must be designed with sensitivity to diverse and intersecting identities, ensuring fairness is contextualized and culturally competent.

Toward Ethical AI: Anti-Racist and Psychologically Informed Approaches

Encouragingly, the growing body of research reviewed in this study signals a shift toward embedding **anti-racist principles** and psychological awareness into AI governance. Integrating models of racism-related stress into psychiatric AI tools, prioritizing participatory design that includes marginalized communities, and mandating transparency and accountability in AI systems represent promising developments.

Effective mitigation of AI bias requires **multidisciplinary collaboration** involving computer scientists, psychologists, ethicists, sociologists, and policymakers. Technical solutions such as bias audits, fairness constraints, and inclusive datasets must be complemented by psychological impact assessments and community engagement to ensure AI systems promote equity and mental well-being.

Limitations and Research Gaps

Despite the comprehensive nature of the data reviewed, several limitations must be acknowledged. Most existing studies focus on Western populations, particularly in the United States and Europe, limiting generalizability to global contexts. There is a notable lack of longitudinal research assessing the long-term psychological effects of AI bias, as well as limited qualitative work exploring lived experiences in depth.

Furthermore, the rapid evolution of AI technology means new forms of bias and impact may be emerging, underscoring the need for ongoing surveillance and adaptive frameworks. This study also relied primarily on secondary data, highlighting the importance of future primary research and participatory methodologies.

Policy and Practice Implications

To mitigate AI bias and its psychological harms, comprehensive policies must be developed that enforce **ethical AI practices** grounded in social justice. This includes:

- Establishing clear standards for bias detection and reporting.
- Requiring transparency in AI system design and decision-making processes.
- Incorporating psychological and social impact assessments in AI development cycles.
- Promoting diversity in AI research teams to reduce blind spots and enhance cultural competence.
- Creating mechanisms for affected communities to participate in governance and have avenues for redress.
- Investing in education and public awareness to build critical understanding of AI's limitations and risks.

Such measures can restore trust and ensure AI serves as a tool for empowerment rather than oppression.

Future Research Directions

This study points to several critical avenues for future research:

- Longitudinal studies exploring the cumulative psychological effects of AI bias.
- Expanding research to include non-Western and underrepresented global populations.
- Developing and testing AI bias mitigation strategies explicitly designed to address psychosocial outcomes.
- Investigating the role of community-led AI governance models.
- Examining how emerging AI technologies (e.g., generative AI, emotion recognition) may create novel forms of bias and psychological impact.

Final Reflections

In conclusion, this research affirms that AI bias is much more than an engineering challenge; it is a pressing **social justice and mental health issue** demanding urgent and sustained attention. Marginalized groups bear a disproportionate burden of AI-induced harms, manifesting not only in economic and social exclusion but also in deep psychological wounds.

Building a future where AI is truly equitable requires **human-centered design**, robust ethical oversight, and a commitment to amplifying marginalized voices. Only through interdisciplinary collaboration and an unwavering focus on justice can AI technologies realize their transformative potential without perpetuating harm.

References

1. Sogancioglu, G., Mosteiro, P., Salah, A. A., Scheepers, F., & Kaya, H. (2024). **Fairness in AI-Based Mental Health: Clinician Perspectives and Bias Mitigation**. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1), 1390–1400. DOI:10.1609/aies.v7i1.31732 AAAI Open Access Journal

2. Rządeczka, M., Sterna, A., Stolińska, J., Kaczyńska, P., & Moskalewicz, M. (2024). **The Efficacy of Conversational Artificial Intelligence in Rectifying the Theory of Mind and Autonomy Biases: Comparative Analysis**. *arXiv preprint*. arXiv

3. Heinz, M. V., Bhattacharya, S., Trudeau, B., Quist, R., Song, S. H., Lee, C. M., Jacobson, N. C. (2023). **Testing Domain Knowledge and Risk of Bias of a Large-Scale General AI Model in Mental Health**. *Digital Health*, 9. SAGE Journals

4. World Health Organization/Europe. (2023). **Artificial intelligence in mental health research: New WHO study on applications and challenges**. *WHO*. World Health Organization

5. ———. (2025). **Artificial intelligence in mental health care: A systematic review of diagnosis, monitoring, and intervention applications**. *Psychological Medicine*, 55, e18. DOI:10.1017/S0033291724003295 Cambridge University Press & Assessment

6. Chen, F., Wang, L., Hong, J., Jiang, J., & Zhou, L. (2023). **Unmasking Bias in AI: A Systematic Review of Bias Detection and Mitigation Strategies in EHR-Based Models**. *arXiv preprint*. arXiv

7. Columbia University et al. (2024). **Patient Perspectives on AI for Mental Health Care: Cross-Sectional Survey Study**. *JMIR Mental Health*, 11. JMIR Mental Health

8. Guntuku, S. C., et al. (2024). **AI fails to detect depression signs in social media posts by Black Americans, study finds**. *Reuters*. Reuters

9. Gabriel, B., Ghassemi, M., et al. (2024). **AI chatbots can detect race, but racial bias reduces response empathy**. *MIT News*. MIT News

10. Bagnis, A., Thayer, J. F., & Mattarozzi, K. (2024). **Racial biases, facial trustworthiness, and resting heart rate variability**. *Cognitive Research: Principles and Implications*, 9, 69. SpringerOpen

11. NIST (2019). **Federal study of top facial recognition algorithms finds 'empirical evidence' of bias**. *The Verge*. The Verge

12. Steed, R., & Caliskan, A. (2021). **A set of distinct facial traits learned by machines is not predictive of appearance bias in the wild**. *AI and Ethics*, 1, 249–260. DOI:10.1007/s43681-020-00035-y SpringerLink

Overview of Source Types

- **Empirical Studies** (peer-reviewed): Sogancioglu et al. (2024); Heinz et al. (2023); Bagnis et al. (2024); Steed & Caliskan (2021).
- **Systematic Reviews & Meta-Analyses**: WHO (2023); *Psychological Medicine* review (2025); Chen et al. (2023).
- **Public Surveys & Media Reports**: *JMIR Mental Health* (2024); Reuters (2024); MIT News (2024).
- **Technical Reports on Facial Recognition Bias**: NIST via *The Verge* (2019).