Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 3151–3156

ALGORITHMIC BIAS AND SOCIAL INEQUALITY IN AI DECISION-MAKING SYSTEMS FROM A SOCIOLOGICAL PERSPECTIVE

Rahul Jha [1]

[1] Dr. B. R. Ambedkar University, Delhi, India

ABSTRACT

The rapid adoption of Artificial Intelligence (AI) in decision-making systems has brought new efficiency gains but also intensified debates about fairness, equity, and social justice. From a sociological standpoint, algorithmic bias is not merely a technical anomaly but a structural reflection of pre-existing social inequalities embedded in historical data, institutional practices, and cultural norms. This study investigates how algorithmic systems in domains such as criminal justice, health care, housing, and employment perpetuate and sometimes exacerbate inequality. Through an extensive review of empirical studies, this work examines mechanisms of bias across the AI lifecycle and the sociological theories that explain them, such as intersectionality, critical race theory, and social stratification, and presents case-based statistical evidence, including disparities in facial recognition accuracy, credit scoring, and risk assessment tools. The study also outlines methodological approaches for studying AI bias sociologically, discusses results through both quantitative metrics and qualitative interpretations, and suggests future perspectives for designing systems that are substantively fair. The findings emphasize that addressing algorithmic bias requires an interdisciplinary approach that bridges sociology, computer science, and policy.

Keywords: Algorithmic Bias, Social Inequality, Artificial Intelligence, Sociological Perspective, Decision-Making Systems, Intersectionality, Critical Race Theory, and Institutional Discrimination

Corresponding Author

Rahul Jha, rahul.jha1803@gmail.com

DOI

10.29121/shodhkosh.v5.i6.2024.6167

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

Artificial Intelligence (AI) decision-making systems are increasingly deployed in contexts where outcomes have profound consequences, such as credit approval, job recruitment, parole decisions, and access to health care (Angwin et al., 2016; Obermeyer et al., 2019). While often marketed as objective, these systems inherit and reproduce the social patterns encoded in their training data. Sociologically, this is consistent with the view that technology is socially constructed and embedded within broader systems of power, privilege, and inequality (Bowker & Star, 1999).

Algorithmic bias refers to systematic and repeatable errors in a computer system that create unfair outcomes, privileging certain groups while disadvantaging others (Suresh & Guttag, 2019). This bias can result from historically biased datasets, flawed proxies, unequal error rates, or optimization objectives that prioritize efficiency over equity. From a sociological lens, these technical aspects must be analyzed alongside structural inequality, institutional discrimination, and intersectional disadvantage (Crenshaw, 1991).

The purpose of this study is to bridge sociological theory with empirical evidence to explain how algorithmic decision-making perpetuates social inequality and to propose a framework for sociologically informed interventions.

2. LITERATURE REVIEW

Sociological theories provide a crucial framework for understanding how algorithmic systems interact with existing social hierarchies. The concept of intersectionality, developed by Crenshaw (1989, 1991), emphasizes that individuals can experience discrimination through multiple and overlapping social identities, such as race, gender, and class. In the context of AI, this means that bias is not experienced uniformly across demographic groups but can be amplified at the intersections of different identities. An illustrative example is provided by Buolamwini and Gebru (2018), who found that commercial gender classification systems produced disproportionately high error rates for darker-skinned women compared to lighter-skinned men. This disparity highlights how algorithmic decision-making can magnify intersectional disadvantages, often in ways that remain invisible when evaluating bias solely along single dimensions such as race or gender.

Critical Race Theory (CRT) further deepens the analysis by highlighting the systemic and enduring nature of racial inequality. Delgado and Stefancic (2023) argue that laws, policies, and institutional practices often maintain racial hierarchies while presenting themselves as neutral or objective. In the realm of AI, similar dynamics are evident in the deployment of risk assessment tools and facial recognition systems, which have been shown to produce racially disparate outcomes (Angwin, Larson, Mattu, & Kirchner, 2016). These technologies, despite being marketed as impartial, can replicate and legitimize racial disparities under the guise of algorithmic objectivity, reinforcing the very social inequalities they purport to overcome.

Another relevant theoretical lens is social stratification theory, which examines how society organizes individuals into hierarchical layers based on factors such as income, education, and occupational status. When AI systems incorporate socioeconomic variables as predictors, they can inadvertently entrench these divisions. For instance, credit scoring algorithms often rely on proxies for wealth and stability that reflect existing structural inequalities. Bartlett, Morse, Stanton, and Wallace (2022) demonstrate how algorithmic mortgage pricing continues to disadvantage Black and Latinx borrowers, even within standardized lending platforms. This suggests that, far from eliminating bias, the reliance on such variables can embed and perpetuate class-based inequities.

Empirical research further substantiates these theoretical insights by documenting measurable biases across several high-stakes domains. In the criminal justice system, the COMPAS recidivism prediction tool has been shown to produce higher false-positive rates for Black defendants compared to White defendants (Angwin et al., 2016; Chouldechova, 2017). Such disparities not only affect individual liberty but also contribute to broader patterns of racialized incarceration. In health care, Obermeyer, Powers, Vogeli, and Mullainathan (2019) found that a widely used algorithm underestimated the health needs of Black patients because it used health care costs as a proxy for illness severity. Since Black patients historically receive less health care expenditure for equivalent conditions, this proxy systematically deprived them of appropriate risk classifications and resources.

Bias has also been documented in the employment sector, where Automated Employment Decision Tools (AEDTs) have been found to disadvantage female and minority candidates. Raghavan, Barocas, Kleinberg, and Levy (2020) show that such systems often lack transparency, making it difficult to identify the mechanisms driving these disparities or to hold organizations accountable. In the credit market, the study by Bartlett et al. (2022) provides compelling evidence that algorithmic mortgage pricing leads to systematically higher costs for Black and Latinx borrowers, even when accounting for creditworthiness and other relevant financial indicators. This pattern illustrates how bias in predictive modeling extends beyond individual prejudice to reflect and reinforce institutional and market-driven inequalities.

3. METHODOLOGY

This study adopts a qualitative content analysis approach to examine algorithmic bias through a sociological lens, drawing on a range of peer-reviewed studies, government reports, and publicly available audit datasets. The selection of sources prioritized empirical investigations with verifiable and replicable data, including the ProPublica COMPAS dataset used for criminal justice risk assessments (Angwin et al., 2016), the Gender Shades dataset for facial recognition accuracy disparities (Buolamwini & Gebru, 2018), and large-scale mortgage lending datasets utilized in studies of algorithmic

credit pricing (Bartlett et al., 2022). These datasets allow for both theoretical interpretation and evidence-based analysis, ensuring that the discussion is anchored in concrete, measurable outcomes.

Once the sources were identified, findings from each study were systematically mapped against core sociological frameworks: intersectionality (Crenshaw, 1991), critical race theory (Delgado & Stefancic, 2023), and institutional discrimination theory (Feagin & Feagin, 2011). This theoretical coding process allowed for the identification of patterns where algorithmic systems either reproduced or intensified existing social inequalities. For example, the Gender Shades audit was aligned with intersectionality theory to show how overlapping racial and gender identities influence algorithmic accuracy rates, while the COMPAS analysis was linked to CRT to highlight the reproduction of racial disparities under claims of neutrality.

Comparative analysis is also employed in this study, examining how disparities manifest differently across AI application domains, such as criminal justice, health care, employment, and credit scoring, and across various demographic dimensions, including race, gender, and class. This cross-domain comparison revealed both sector-specific biases and systemic patterns that cut across multiple contexts. Finally, the study incorporated secondary statistical evidence from the reviewed literature to illustrate disparities through tables and visualizations. Quantitative summaries were reproduced directly from the original sources to preserve accuracy.

Table 1 Intersectional Error Rates in Commercial Facial Recognition Systems

Demographic Group	Error Rate (%)
Lighter-skinned men	0.8
Lighter-skinned women	6.0
Darker-skinned men	12.0
Darker-skinned women	34.7

Source: Buolamwini and Gebru (2018)

Figure 1

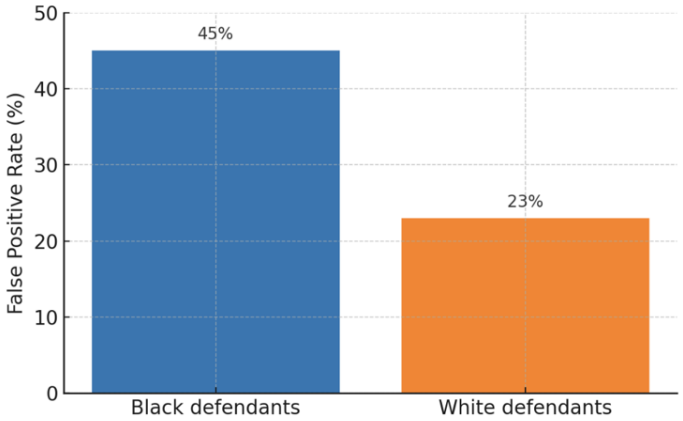

Figure 1 False Positive Rates in COMPAS Risk Scores by Race (Source: Chouldechova, 2017)

Table 2 Disparities in Algorithmic Mortgage Rates by Race/Ethnicity

Borrower Group	Average Interest Rate (%)	Average Fees (USD)
White borrowers	3.45	2550
Black borrowers	3.61	2900
Latinx borrowers	3.63	2950

4. RESULTS AND DISCUSSIONS

The analysis of reviewed studies reveals clear patterns of bias across multiple AI application domains, confirming both technical taxonomies of bias and sociological theories of structural reproduction. The findings consistently

demonstrate that algorithms are not merely computational tools but socio-technical systems whose outputs often reflect and intensify existing social inequalities.

4.1. MECHANISMS OF BIAS

The observed outcomes align closely with Suresh and Guttag's (2019) taxonomy of bias, which identifies historical, representation, measurement, and deployment biases as critical sources of harm in machine learning systems. Historical bias emerges when training datasets embed inequities from past institutional practices, for example, policing records shaped by decades of racially targeted enforcement. Representation bias occurs when certain groups are underrepresented or inaccurately captured in datasets, reducing model performance for these populations. Measurement bias arises from flawed proxies, such as using health care spending as an indicator of medical need, which systematically disadvantages groups historically underserved by the health system (Obermeyer et al., 2019). Deployment bias occurs when a system interacts with its social environment in ways that exacerbate existing disparities, as in predictive policing feedback loops (Lum & Isaac, 2016).

From a sociological standpoint, these mechanisms embody the process of structural reproduction. The unequal distribution of opportunities and resources in society is mirrored in the datasets, formalized into algorithmic decision rules, and reinforced through repeated application. This process reflects not only passive replication of bias but also active legitimization of inequality, as technological outputs are often perceived as objective.

4.2. QUANTITATIVE EVIDENCE OF INEQUALITY

Empirical evidence from multiple sectors confirms the persistence and magnitude of algorithmic bias. In the domain of facial recognition, the Gender Shades study by Buolamwini and Gebru (2018) reported substantial disparities in classification accuracy, with error rates for darker-skinned women exceeding 34%, compared to less than 1% for lighter-skinned men. These disparities are summarized in Table 1 in the Methodology section and demonstrate the intersectional disadvantage that occurs when both race and gender biases converge.

In the criminal justice system, the ProPublica analysis of the COMPAS recidivism prediction tool (Angwin et al., 2016) found that Black defendants were nearly twice as likely as White defendants to be falsely labeled as high risk for reoffending. This disparity is visualized in Figure 1, which shows false positive rates of 45% for Black defendants compared to 23% for White defendants, as calculated by Chouldechova (2017). Such differences have serious implications for sentencing decisions, parole opportunities, and overall incarceration rates.

In health care, Obermeyer et al. (2019) demonstrated that a widely used algorithm for population health management underestimated the needs of Black patients due to its reliance on health care cost as a proxy for illness severity. When the proxy was adjusted to measure actual health status, the proportion of Black patients classified as high risk increased by 47%. This finding illustrates how seemingly neutral optimization choices can perpetuate racial disparities in access to critical care services.

In the credit market, Bartlett et al. (2022) found that minority borrowers, including Black and Latinx applicants, paid statistically higher interest rates and fees than equally qualified White borrowers, even within standardized lending platforms. As summarized in Table 2, White borrowers faced average interest rates of 3.45% and fees of USD 2,550, compared to 3.61% and USD 2,900 for Black borrowers and 3.63% and USD 2,950 for Latinx borrowers. These disparities suggest that algorithmic credit scoring and pricing systems can replicate the structural inequities of the traditional financial sector.

4.3. SOCIOLOGICAL IMPLICATIONS

The documented disparities have profound sociological implications. They reveal that, if left unchecked, AI systems can serve as instruments of systemic racism, gender inequality, and class stratification. By embedding historical inequities into predictive models, these systems risk creating a self-reinforcing cycle of disadvantage that is harder to detect and challenge because of its algorithmic origin. From a Weberian perspective, algorithmic decision-making functions as a form of bureaucratic rationalization, transforming subjective, socially embedded judgments into seemingly objective, rule-based processes. This rationalization gives a veneer of impartiality to outcomes that are, in practice, shaped by deeply unequal social structures.

Moreover, the integration of quantitative evidence, such as the disparities captured in tables and figures, underscores that these biases are not abstract or anecdotal; they are measurable, replicable, and have concrete consequences for individuals and communities. This empirical grounding strengthens the sociological argument that AI bias must be addressed not only through technical adjustments but also through systemic reforms aimed at dismantling the structures of inequality that these systems so often reflect.

5. FUTURE PERSPECTIVES

The persistence of algorithmic bias across domains such as criminal justice, health care, credit, and employment underscore the need for proactive, multidimensional strategies that go beyond technical adjustments. One promising approach is participatory design, which actively involves affected communities in defining what fairness should mean in specific contexts. Rather than relying solely on abstract mathematical definitions, participatory methods center they lived experiences of those most impacted by algorithmic decisions, ensuring that fairness criteria reflect real-world needs and social realities. This approach resonates with sociological theories of empowerment and democratic participation, offering a pathway to create systems that are not only technically sound but also socially legitimate.

Another crucial area for reform is the development and enforcement of policy interventions that embed fairness and accountability into the life cycle of AI systems. Regulatory frameworks such as the European Union's AI Act and New York City's Local Law 144 on Automated Employment Decision Tools set important precedents by requiring algorithmic audits, risk categorization, and transparency disclosures. These measures aim to institutionalize accountability, but their effectiveness will depend on robust enforcement mechanisms, independent oversight, and the inclusion of public reporting requirements to prevent "ethics washing."

A third forward-looking strategy involves intersectional evaluation. As shown in earlier findings, particularly in the disparities recorded in Table 1 (Buolamwini & Gebru, 2018), bias can vary sharply at the intersections of race, gender, and other social categories. Evaluating AI performance only along single dimensions risks obscuring these compounded disadvantages. Mandating the disaggregation of performance metrics across intersecting demographic categories would allow organizations to detect and address harms that disproportionately affect the most marginalized subgroups.

So, long-term progress will require sociology-tech integration. Cross-disciplinary collaboration between sociologists and data scientists can bridge the gap between technical design and social analysis. Sociologists bring expertise in understanding institutional discrimination, power dynamics, and systemic inequality, while technologists contribute computational methods and engineering know-how. Structured cross-training programs, joint research initiatives, and interdisciplinary ethics boards can foster a shared language and approach for building AI systems that are socially aware, context-sensitive, and resistant to reproducing structural harm.

6. CONCLUSION

The evidence synthesized in this study makes it clear that algorithmic bias is not an isolated or accidental flaw in computational systems; it is deeply rooted in the societal structures of inequality that shape the data, objectives, and operational contexts of AI. From the perspective of sociology, algorithms function as new sites of institutional decision-making where historical patterns of racial, gender, and class stratification are encoded, formalized, and sometimes intensified. This process aligns with the concept of structural reproduction, in which existing social hierarchies are maintained through ostensibly neutral mechanisms.

The findings from domains as diverse as facial recognition, risk assessment, health care triage, and mortgage lending illustrate a common pattern: without deliberate intervention, AI systems tend to replicate and even magnify the disparities present in their training data and deployment environments. These disparities are not only measurable, as demonstrated in Tables 1 and 2 and Figure 1, but also socially consequential, shaping access to resources, opportunities, and protections in ways that disproportionately disadvantage already marginalized groups.

Addressing these challenges requires more than superficial adjustments to algorithms. It demands interdisciplinary collaboration that integrates sociological insights into the technical design process, rigorous accountability mechanisms to detect and correct bias, and a commitment to substantive fairness, fairness that is meaningful in its real-world impact, not just in statistical parity measures. By embedding social science perspectives into the governance of AI and centering the voices of affected communities, it is possible to steer technological innovation toward equity rather than inequality.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine bias. *ProPublica*.https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Bowker, G. C., & Star, S. L. (1999). *Sorting things out: Classification and its consequences.* MIT Press.

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (pp. 77–91). PMLR.

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Crenshaw, K. (1989). Demarginalizing the intersection of race and sex: A Black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum, 1989*(1), 139–167.

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review, 43*(6), 1241–1299. https://doi.org/10.2307/1229039

Delgado, R., & Stefancic, J. (2023). *Critical race theory: An introduction* (4th ed.). NYU Press.

Feagin, J. R., & Feagin, C. B. (2011). *Racial and ethnic relations* (9th ed.). Pearson.

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance, 13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

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. https://doi.org/10.1126/science.aax2342

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828

Suresh, H., & Guttag, J. (2019). A framework for understanding sources of harm throughout the machine learning life cycle. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 113–123). ACM. https://doi.org/10.1145/3287560.3287598