Siddhanta's International Journal of Advanced Research in Arts & Humanities
Vol. 3, No. 1, September – October 2025

Algorithmic Bias and Social Stratification: How AI Shapes Inequality in Digital Societies

Parhlad Singh Ahluwalia, School of Management and Business Studies, Jamia Hamdard, New Delhi, India
Email : editor@shodhprakashan.in

Abstract

This paper examines the complex relationship between algorithmic bias and social stratification in contemporary digital societies. As artificial intelligence systems increasingly mediate access to employment, credit, healthcare, and education, their embedded biases risk perpetuating and amplifying existing social inequalities. Through analysis of case studies across multiple domains, this research demonstrates how algorithmic decision-making systems can systematically disadvantage marginalized groups, creating new forms of digital stratification. The paper proposes a framework for understanding algorithmic bias as both a technical and social phenomenon, arguing that addressing these issues requires interdisciplinary approaches combining technical solutions with policy interventions and social awareness. Findings suggest that without deliberate intervention, AI systems may entrench existing power structures while creating novel forms of discrimination that are harder to detect and challenge.

Keywords: algorithmic bias, social stratification, artificial intelligence, digital inequality, machine learning, discrimination, social justice, technology ethics

Introduction

The proliferation of algorithmic decision-making systems across various sectors of society has fundamentally altered how individuals access opportunities and resources. From hiring algorithms that screen job applications to credit scoring systems that determine loan eligibility, artificial intelligence increasingly serves as a gatekeeper for social and economic mobility (O'Neil, 2016). While these systems promise efficiency and objectivity, mounting evidence suggests they often perpetuate and amplify existing social biases, creating new forms of digital inequality.

Social stratification, traditionally understood as the hierarchical arrangement of individuals and groups based on factors such as wealth, power, and prestige, has found new expression in the digital age. Algorithmic systems, rather than serving as neutral arbiters, have become active participants in the reproduction of social hierarchies. This phenomenon occurs through

various mechanisms, including biased training data, flawed algorithmic design, and the deployment of AI systems within existing structures of inequality.

The significance of this issue extends beyond individual instances of unfair treatment. Algorithmic bias operates at scale, potentially affecting millions of decisions simultaneously and creating systemic patterns of discrimination that may be difficult to detect or challenge. Unlike human bias, which is localized and inconsistent, algorithmic bias can be replicated indefinitely, creating what Cathy O'Neil terms "weapons of math destruction" that punish the poor and marginalized while reinforcing privilege for the advantaged.

This paper argues that algorithmic bias represents a critical mechanism through which social stratification is maintained and extended in digital societies. By examining the technical, social, and policy dimensions of this phenomenon, we can better understand how AI systems shape inequality and develop more effective strategies for creating fairer algorithmic systems.

Literature Review

Theoretical Foundations of Algorithmic Bias

The concept of algorithmic bias has evolved from early concerns about computer fairness to a sophisticated understanding of how automated systems can perpetuate discrimination. Barocas and Selbst (2016) provided foundational work by categorizing different types of bias that can emerge in machine learning systems, distinguishing between bias in training data, algorithmic processing, and outcome interpretation. Their framework demonstrates how bias can enter systems at multiple stages, making it a persistent challenge rather than a simple technical problem to solve.

Building on this foundation, Friedman and Nissenbaum (1996) introduced the concept of "bias in computer systems," arguing that technological artifacts are not neutral but embody the values and assumptions of their creators. This perspective challenges the notion that algorithms are objective, highlighting how human judgments and social contexts inevitably shape automated decision-making processes.

Mechanisms of Digital Stratification

Recent scholarship has identified several key mechanisms through which algorithmic systems contribute to social stratification. Eubanks (2018) documented how automated decision-making systems in social services create "digital poorhouses" that surveil and punish low-income individuals while providing conveniences for the wealthy. Her ethnographic work reveals how algorithmic systems can institutionalize class-based discrimination under the guise of efficient resource allocation.

Noble (2018) extended this analysis to examine how search algorithms perpetuate racial and gender stereotypes, demonstrating how seemingly neutral information retrieval systems can reinforce harmful social hierarchies. Her work on "algorithms of oppression" shows how commercial interests and existing social biases become embedded in systems that shape public understanding and opportunity.

Empirical Evidence of Algorithmic Discrimination

Empirical studies have documented algorithmic bias across numerous domains. In criminal justice, Angwin et al. (2016) found that risk assessment algorithms used in sentencing decisions exhibited significant racial bias, incorrectly flagging Black defendants as high-risk at nearly twice the rate of white defendants. This finding sparked widespread debate about the use of algorithmic tools in the justice system and highlighted the potential for AI to perpetuate racial discrimination.

In employment, Dastin (2018) reported that Amazon scrapped an AI recruiting tool after discovering it systematically downgraded resumes from women, particularly for technical positions. The system had learned to associate male-dominated hiring patterns with success, effectively automating gender discrimination in the hiring process.

Healthcare presents another domain where algorithmic bias has been documented. Obermeyer et al. (2019) found that a widely-used healthcare algorithm exhibited significant racial bias, underestimating the health needs of Black patients compared to white patients with similar health conditions. This bias occurred because the algorithm used healthcare spending as a proxy for health needs, failing to account for systemic inequalities in healthcare access.

Methodology

This paper employs a mixed-methods approach combining literature review, case study analysis, and theoretical synthesis. The research draws on peer-reviewed academic literature, investigative journalism, and technical reports from industry and civil society organizations. Case studies were selected to represent different domains where algorithmic decision-making has significant social impact, including criminal justice, employment, finance, healthcare, and education.

The analysis framework considers three levels of investigation: individual cases of bias, systemic patterns across domains, and broader implications for social stratification. This multi-level approach allows for both detailed examination of specific instances of algorithmic bias and broader theoretical insights about the relationship between AI and social inequality.

Analysis and Findings

Case Study 1: Criminal Justice Risk Assessment

Risk assessment algorithms in criminal justice exemplify how algorithmic bias can perpetuate racial inequality. The Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) system, widely used across the United States, attempts to predict the likelihood of recidivism among defendants. However, analysis by ProPublica revealed significant racial disparities in the system's predictions (Angwin et al., 2016).

The algorithm's bias manifests in several ways. Black defendants are nearly twice as likely to be incorrectly classified as high-risk compared to white defendants, while white defendants are more likely to be incorrectly classified as low-risk. These disparities occur even when controlling for other factors such as age and criminal history. The consequences are severe, as risk scores influence decisions about bail, sentencing, and parole, potentially extending incarceration for Black defendants based on biased predictions.

The technical roots of this bias trace to training data that reflects historical patterns of discriminatory policing and sentencing. Areas with higher police presence generate more arrests and convictions, creating feedback loops that reinforce geographic and racial targeting. The algorithm learns these patterns as predictive signals, effectively automating discriminatory practices while obscuring them behind a veneer of scientific objectivity.

Case Study 2: Employment Screening Algorithms

Automated hiring systems represent another domain where algorithmic bias significantly impacts social stratification. These systems promise to reduce human bias in hiring decisions by focusing on objective qualifications and skills. However, evidence suggests they often perpetuate existing inequalities while creating new forms of discrimination.

Amazon's scrapped recruiting algorithm provides a prominent example. The system was trained on resumes submitted to the company over a ten-year period, during which hiring was predominantly male, particularly in technical roles. The algorithm learned to associate this pattern with successful candidates, systematically downgrading resumes that included words associated with women, such as "women's chess club captain" or graduates of all-women's colleges (Dastin, 2018).

This case illustrates how historical bias becomes embedded in algorithmic systems through training data. The algorithm's "objectivity" merely automated existing patterns of discrimination, making them appear neutral and scientific. Moreover, because the bias was embedded in the system's learned associations rather than explicit rules, it was difficult to detect and correct without extensive analysis.

Case Study 3: Healthcare Resource Allocation

Healthcare algorithms demonstrate how bias can emerge from seemingly neutral proxy variables. Obermeyer et al. (2019) examined a widely-used algorithm that determined which patients received additional healthcare resources and support. The system was designed to identify patients with the greatest health needs, but analysis revealed significant racial bias in its predictions.

The algorithm used healthcare spending as a proxy for health needs, reasoning that patients with greater medical needs would incur higher costs. However, this approach failed to account for systemic inequalities in healthcare access. Black patients historically receive less care due to factors including provider bias, geographic access barriers, and insurance coverage disparities. Consequently, Black patients with the same health conditions as white patients typically generate lower healthcare costs, leading the algorithm to underestimate their needs.

This case highlights how algorithmic bias can emerge from structural inequalities rather than explicit discrimination. The bias was not intentional but resulted from using proxy variables that reflect existing disparities. The algorithm's apparent neutrality masked its role in perpetuating healthcare inequality, potentially denying additional support to patients who needed it most.

Table 1: Domains of Algorithmic Bias and Their Social Impact

Domain	Algorithm Type	Bias Mechanism	Affected Groups	Social Impact
Criminal Justice	Risk Assessment	Historical discrimination in training data	Racial minorities	Extended incarceration, limited opportunities
Employment	Resume screening, interview selection	Male-dominated training data	Women, minorities	Reduced job opportunities, wage gaps
Healthcare	Resource allocation, diagnosis	Spending-based proxies	Racial minorities, low-income	Inadequate care, health disparities
Finance	Credit scoring, loan approval	Traditional credit markers	Minorities, young adults	Limited access to credit, economic mobility
Education	Admissions, placement	Socioeconomic proxies	Low-income, first-generation students	Educational stratification
Housing	Rental screening, mortgage approval	Geographic and demographic proxies	Minorities, low-income	Residential segregation

Systemic Patterns and Feedback Loops

Analysis across domains reveals several systemic patterns in how algorithmic bias perpetuates social stratification. First, many biased algorithms rely on proxy variables that correlate with protected characteristics but appear neutral. Credit scores, zip codes, educational credentials, and even language patterns can serve as proxies for race, gender, or socioeconomic status, allowing discrimination to occur without explicit use of protected categories.

Second, algorithmic systems often create feedback loops that reinforce existing inequalities. Predictive policing algorithms direct police attention to areas with historically high arrest rates, potentially increasing surveillance and arrests in these communities regardless of actual crime rates. Similarly, hiring algorithms trained on past hiring decisions may perpetuate existing workplace demographics by learning to associate certain characteristics with "successful" candidates.

Third, the opacity of many algorithmic systems makes bias difficult to detect and challenge. Complex machine learning models often function as "black boxes" whose decision-making processes are not easily interpretable. This opacity protects algorithmic systems from scrutiny and makes it difficult for affected individuals to understand or contest biased decisions.

Intersectionality and Compound Disadvantage

Algorithmic bias often exhibits intersectional effects, where individuals with multiple marginalized identities face compound disadvantage. For example, Black women may face both racial and gender bias in hiring algorithms, while low-income minorities may be disadvantaged across multiple algorithmic systems simultaneously. These intersectional effects can create particularly severe forms of digital exclusion that are difficult to address through single-axis bias mitigation strategies.

The compound nature of algorithmic disadvantage means that individuals may face a cascade of biased decisions across multiple life domains. A biased criminal justice algorithm may lead to longer incarceration, which in turn affects employment prospects when hiring algorithms flag criminal history. Poor employment outcomes may then impact credit scores, affecting housing and financial opportunities. This cascade effect demonstrates how algorithmic bias can create systematic patterns of exclusion that reinforce social stratification.

Discussion

Algorithmic Bias as a Social Phenomenon

The evidence presented demonstrates that algorithmic bias is not merely a technical problem but a social phenomenon that reflects and amplifies existing power structures. Biased

algorithms emerge from biased data, biased design choices, and deployment within biased social systems. Technical solutions alone are insufficient to address these issues because they fail to address the underlying social conditions that produce bias.

This understanding challenges the common framing of algorithmic bias as an unfortunate side effect of technological progress that can be solved through better engineering. Instead, algorithmic bias represents a systematic feature of how AI systems interact with existing social hierarchies. Addressing these issues requires recognizing algorithms as sociotechnical systems embedded within broader patterns of inequality.

The Myth of Algorithmic Objectivity

One of the most significant barriers to addressing algorithmic bias is the persistent myth of algorithmic objectivity. This myth holds that automated systems are inherently more fair and objective than human decision-makers because they are not subject to conscious bias or emotional influence. However, this perspective ignores how human judgments, biases, and social contexts are embedded within algorithmic systems at every stage of development and deployment.

The myth of algorithmic objectivity serves to legitimize biased decisions by attributing them to neutral, scientific processes. When algorithmic systems produce discriminatory outcomes, these are often dismissed as unfortunate errors rather than recognized as systematic patterns that require structural intervention. This legitimization makes algorithmic bias particularly pernicious because it obscures discrimination behind a veneer of technological neutrality.

Implications for Social Justice

The findings have significant implications for social justice advocacy and policy intervention. Traditional approaches to addressing discrimination, which focus on intentional bias and disparate treatment, may be insufficient for addressing algorithmic discrimination. Algorithmic bias often operates through disparate impact rather than disparate treatment, creating discriminatory outcomes through seemingly neutral processes.

Moreover, the scale and scope of algorithmic decision-making mean that biased systems can affect far more people than individual instances of human discrimination. A single biased algorithm may influence thousands or millions of decisions, creating systematic patterns of disadvantage that can persist indefinitely unless actively addressed.

Toward Algorithmic Justice

Addressing algorithmic bias requires a comprehensive approach that combines technical, legal, and social interventions. Technical approaches include developing bias detection and mitigation tools, creating more diverse and representative training datasets, and improving

algorithmic transparency and interpretability. However, these technical solutions must be complemented by broader social changes.

Legal interventions may include extending civil rights protections to cover algorithmic discrimination, requiring algorithmic impact assessments for systems used in high-stakes decisions, and creating enforcement mechanisms for algorithmic fairness. Policy interventions might focus on regulating the use of certain proxy variables, requiring transparency in algorithmic decision-making, and ensuring meaningful human oversight of automated systems.

Social interventions involve changing the broader context within which algorithmic systems operate. This includes addressing underlying social inequalities that produce biased training data, diversifying the technology workforce to bring different perspectives to algorithmic design, and building public awareness of algorithmic bias issues.

Table 2: Strategies for Addressing Algorithmic Bias

Level	Strategy	Examples	Limitations
Technical	Bias detection and mitigation	Fairness metrics, debiasing algorithms	May not address root causes
Individual	User education and rights	Algorithm literacy, appeal processes	Limited individual power
Organizational	Internal governance	Ethics boards, bias auditing	Voluntary compliance issues
Legal	Regulatory frameworks	Anti-discrimination law, transparency requirements	Enforcement challenges
Social	Structural change	Addressing underlying inequality, workforce diversity	Long-term, complex process

Limitations and Future Research

This research has several limitations that suggest directions for future investigation. First, the rapidly evolving nature of AI technology means that new forms of algorithmic bias may emerge faster than they can be studied and addressed. Future research should develop more proactive approaches to identifying potential bias in emerging AI applications.

Second, most current research on algorithmic bias focuses on Western, developed country contexts. The global deployment of AI systems raises questions about how algorithmic bias manifests in different cultural, legal, and social contexts. Cross-cultural research on

algorithmic bias could provide valuable insights for developing more universally applicable solutions.

Third, the intersectional nature of algorithmic bias requires more sophisticated analytical frameworks that can capture how multiple forms of bias interact and compound. Future research should develop better methods for studying and addressing intersectional algorithmic discrimination.

Finally, more research is needed on the effectiveness of different interventions for addressing algorithmic bias. While many technical and policy solutions have been proposed, empirical evidence on their real-world effectiveness remains limited.

Conclusion

This analysis demonstrates that algorithmic bias represents a significant mechanism through which social stratification is maintained and extended in digital societies. Far from being neutral arbiters, AI systems actively participate in the reproduction of social hierarchies through biased training data, flawed design choices, and deployment within existing structures of inequality.

The evidence reveals that algorithmic bias operates across multiple domains, affecting access to employment, credit, healthcare, education, and justice. These biased systems create feedback loops that reinforce existing inequalities while generating new forms of digital discrimination that are often harder to detect and challenge than traditional forms of bias.

Addressing algorithmic bias requires recognizing it as a social phenomenon rather than merely a technical problem. Solutions must combine technical innovations with legal protections, policy interventions, and broader social changes. The goal should not be merely to eliminate bias from individual algorithms but to create algorithmic systems that actively promote fairness and social justice.

As AI systems become increasingly central to social and economic life, the stakes of this work continue to grow. Without deliberate intervention, algorithmic bias risks creating a digital caste system where technological systems systematically advantage some groups while disadvantaging others. Preventing this outcome requires sustained effort from technologists, policymakers, civil society organizations, and affected communities working together to ensure that AI systems serve all members of society fairly.

The path forward involves not just fixing biased algorithms but reimagining how AI systems can be designed and deployed to promote rather than undermine social equality. This represents both a technical challenge and a social opportunity to create digital systems that

reflect our highest aspirations for justice and fairness rather than our historical patterns of discrimination and exclusion.

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
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671-732.
- Dastin, J. (2018, October 10). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
- Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.
- Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3), 330-347.
- Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. New York University Press.
- 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.
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown Publishers.