International Journal of Education and Humanities
ISSN: 2770-6702 | Vol. 20, No. 2, 2025

Algorithmic Justice: Can AI Mitigate or Exacerbate Bias in Criminal Sentencing?

Yiqiang Gao*

East China University of Political Science and Law, Shanghai 201620, China

* **Corresponding author:** Yiqiang Gao (Email: admin@csf.ac.cn)

Abstract: This paper explores AI's dual role in mitigating or exacerbating bias in criminal sentencing. As AI-driven tools increasingly integrate into global criminal justice systems, algorithmic justice has become a critical concern for legal practitioners, policymakers, and technologists. The study examines how AI can reduce bias through data-driven objectivity, enhanced consistency in decision-making, and comprehensive factor analysis that addresses limitations in human cognitive processing. Simultaneously, it investigates how flawed training data, problematic algorithm design, and interpretability gaps can amplify existing inequalities within judicial systems. Through comparative case studies of successful and problematic AI implementations, the research identifies key factors influencing AI's impact on sentencing equity. It proposes practical strategies including rigorous data preprocessing protocols, systematic algorithmic auditing, and the establishment of robust legal and ethical frameworks. Findings reveal that AI's impact on criminal sentencing is neither inherently beneficial nor harmful but is shaped by data quality, algorithm fairness, and governance structures, offering valuable guidance for responsible AI implementation in criminal justice.

Keywords: Algorithmic Justice, Artificial Intelligence, Criminal Sentencing, Bias Mitigation, Algorithmic Bias, Criminal Justice, Risk Assessment Algorithms, Legal Ethics.

1. Introduction

1.1. Background of AI in Criminal Sentencing

Global criminal justice systems are increasingly turning to artificial intelligence to address mounting caseloads, resource constraints, and growing public demands for both efficiency and fairness. AI technologies offer unique capabilities to process vast quantities of case data, identify complex patterns, and generate consistent recommendations that assist judges in making more informed sentencing decisions. Unlike human decision-makers-who are prone to cognitive biases, fatigue, and information overload-AI systems can systematically evaluate multiple variables simultaneously, including crime severity, aggravating and mitigating circumstances, and defendant background characteristics.

A key driver behind AI adoption is the pursuit of greater consistency in sentencing outcomes. The fundamental judicial principle of "like cases treated alike" has long been challenged by empirical studies documenting significant disparities in sentences for similar offenses. These disparities are influenced by factors ranging from judge-specific characteristics (such as judicial philosophy or demographic background) to contextual variables (including geographic location and even the time of day or week). AI promises to reduce such variability by applying uniform analytical frameworks across cases, potentially enhancing both fairness and public trust in judicial outcomes.

1.2. Research Objectives

This research aims to explore the complex and multifaceted role of AI in shaping bias in criminal sentencing. Its primary objectives include examining how biased training data-reflecting historical or systemic inequalities-can lead AI systems to perpetuate or amplify existing disparities. The study evaluates AI performance across diverse legal, cultural, and socioeconomic contexts to identify contextual factors that influence algorithmic fairness. It investigates the dynamic interaction between AI-generated recommendations and human judicial decision-making, analyzing how judges interpret, utilize, and potentially override algorithmic outputs. Finally, the research identifies best practices for developing and deploying AI tools that maximize bias mitigation while minimizing risks of exacerbating inequalities.

1.3. Significance of the Study

The practical significance of this research lies in its potential to guide evidence-based implementation of AI in criminal justice systems. By identifying conditions under which AI reduces bias, the study justifies targeted investment in responsible AI development. Conversely, by highlighting risks and failure modes, it informs the creation of ethical guidelines and regulatory frameworks to prevent algorithmic harm. Academically, the research contributes to interdisciplinary discourse at the intersection of law, computer science, and social justice, deepening understanding of how technological tools interact with systemic inequalities. It opens new avenues for research in algorithmic fairness, legal ethics, and the sociology of technology, while providing a critical analysis of the promises and pitfalls of technological solutions to complex social problems.

2. Understanding AI and Criminal Sentencing

2.1. Basics of AI Algorithms in the Legal Field

AI algorithms used in criminal sentencing primarily utilize three machine learning approaches: supervised learning, unsupervised learning, and deep learning. Supervised learning models-most common in sentencing applications-are trained on labeled datasets where past cases are paired with

their actual sentencing outcomes. These models learn to predict sentences for new cases by identifying patterns in input variables (such as crime details and defendant characteristics) that correlate with historical sentencing decisions. This approach powers most risk assessment tools that predict recidivism or flight risk.

Unsupervised learning algorithms analyze unlabeled data-such as court transcripts, police reports, or probation records-to identify hidden patterns and group similar cases without prior outcome knowledge. These models help detect anomalies in sentencing practices, such as clusters of unusually harsh or lenient sentences for specific case types. Deep learning, using multi-layer neural networks, processes diverse data types including text, images, and numerical data, enabling more nuanced analysis of complex cases involving multiple evidentiary sources.

The performance of these algorithms is fundamentally dependent on training data quality. If datasets reflect historical biases-such as over-representation of certain ethnic groups due to discriminatory policing practices-algorithms will learn and replicate these biases, potentially producing unfair sentencing recommendations despite technical sophistication.

2.2. Current Applications of AI in Criminal Sentencing

Prominent AI applications in criminal sentencing include risk assessment tools like COMPAS (Correctional Offender Management Profiling for Alternative Sanctions), which generates recidivism risk scores using variables including criminal history, age, employment status, and residential stability. These scores inform decisions ranging from pretrial detention to sentence length and supervision intensity. Case matching systems use AI to identify precedential cases, comparing current cases to thousands of historical records to suggest relevant sentencing ranges and ensure consistency with past decisions.

AI also automates pre-sentencing report preparation by analyzing police reports, victim statements, and probation assessments to compile comprehensive defendant profiles, saving judicial resources while ensuring systematic consideration of relevant factors. Additionally, algorithmic monitoring systems analyze sentencing patterns across judges and jurisdictions to identify potential disparities, enabling oversight bodies to investigate and address inconsistent practices.

2.3. The Concept of Criminal Sentencing and Its Ideal Standards

Criminal sentencing involves determining appropriate punishment for convicted offenders, balancing multiple objectives including retribution (just deserts), deterrence (preventing future crime), rehabilitation (reform), and incapacitation (protecting society). Ideal sentencing standards emphasize justice through unbiased treatment of similarly situated offenders, regardless of irrelevant characteristics like race, gender, or socioeconomic status.

Proportionality requires punishment severity to match offense seriousness, with consideration of both harm caused and offender culpability. Rehabilitation-orientation focuses on addressing root causes of criminal behavior through targeted interventions. Transparency demands clear, reasoned explanations for sentencing decisions to maintain public trust. These standards provide a normative framework against which to evaluate both human and algorithmic sentencing practices.

3. The Potential of AI to Mitigate Bias in Criminal Sentencing

3.1. Objectivity in Data-Driven Decision Making

AI introduces greater objectivity by reducing reliance on human cognitive biases that influence sentencing. Psychological research shows judges are unconsciously affected by extraneous factors including defendant appearance, mannerisms, and socioeconomic cues (Dressler & Michaels, 2009)[5]. AI systems, by contrast, focus exclusively on predefined legal and factual factors, minimizing influence from irrelevant characteristics.

A study by Berger (2017)[4] demonstrated this potential, documenting an 8-percentage-point reduction in racial sentencing disparities following AI implementation in a mid-sized jurisdiction. The AI system consistently prioritized legally relevant factors like offense severity and prior convictions while filtering out racial identifiers and correlated variables. Importantly, AI enhances transparency by providing auditable explanations of how specific factors contribute to recommendations, enabling stakeholders to identify and correct residual biases in algorithmic decision-making.

3.2. Enhanced Consistency in Sentencing

AI addresses problematic variability in human judicial decisions. A 2015[3] American Bar Association report found sentence lengths for similar burglary offenses varied by up to 50% across different judges, reflecting individual judicial philosophies and implicit biases. AI systems reduce such variability by applying consistent analytical frameworks to all cases.

Kleinberg et al. (2018)[6] documented a significant reduction in sentencing inconsistency after AI implementation in drug offense cases, with standard deviation in sentence lengths decreasing from 18 months to 10 months. AI ensures consistent application of legal reforms and sentencing guidelines across jurisdictions, preventing geographical disparities where similar cases receive dramatically different outcomes based on location. This consistency provides greater predictability for defendants, victims, and legal professionals, enhancing perceptions of fairness in the criminal justice system.

3.3. Comprehensive Consideration of Factors

AI overcomes human cognitive limitations by systematically processing vast amounts of information, ensuring relevant factors are not overlooked. Judges facing heavy caseloads may inadvertently prioritize certain factors while neglecting others, such as missing mitigating circumstances in complex cases. AI systems simultaneously analyze structured data (criminal records, offense details) and unstructured data (social work reports, psychological evaluations) to develop comprehensive case assessments.

For example, AI can integrate information about a defendant's traumatic background, employment history, and participation in treatment programs-factors that might be overlooked in human decision-making-when recommending sentences that balance punishment with rehabilitation potential. Advanced algorithms identify nuanced interactions

between variables, such as how prior convictions' predictive value varies by age at offense and access to community support, enabling more contextually appropriate sentencing recommendations.

4. How AI Might Exacerbate Bias in Criminal Sentencing

4.1. Biased Data Input

The adage "garbage in, garbage out" applies acutely to AI sentencing tools, as biased training data perpetuates and amplifies existing inequalities. Historical criminal justice data reflects systemic biases in law enforcement and adjudication. Alexander (2010)[1] documented how African-American communities face disproportionate policing, leading to over-representation in arrest and conviction records. When trained on such data, AI systems learn to associate African-American defendants with higher risk, producing biased recommendations even when controlling for legal factors.

Data quality issues compound these problems. Hispanic defendants often have minor offenses mislabeled as serious due to language barriers or cultural misunderstandings, skewing algorithmic perceptions of recidivism risk. Geographic bias occurs when urban-centric data fails to reflect rural crime patterns and social conditions, leading to inaccurate assessments of rural defendants. Under-representation of marginalized groups like Native Americans in training data results in poorly calibrated recommendations that fail to account for unique cultural and historical contexts.

4.2. Algorithm Design Flaws

Even with relatively unbiased data, algorithm design choices can introduce or amplify bias. Risk assessment tools like COMPAS have faced criticism for over-weighting variables such as employment and residential stability—factors heavily influenced by systemic inequalities (Angwin et al., 2016)[2]. This design flaw leads to disproportionate labeling of Black defendants as high-risk, as historical discrimination creates barriers to stable employment and housing.

Algorithms often fail to account for differential factor impacts across demographic groups. For example, relying heavily on prior arrest records without adjusting for biased policing practices penalizes groups targeted by law enforcement. Gender bias emerges in tools that over-emphasize male gender as a risk factor while underweighting offense-specific characteristics. Algorithms optimized solely for predictive accuracy may exploit proxies for protected characteristics, such as using neighborhood data that correlates with race due to historical redlining rather than actual risk factors.

4.3. Lack of Interpretability and Accountability

The "black box" nature of complex AI models-particularly deep learning systems-obscures how sentencing recommendations are generated, impeding bias detection and correction. In one European case, a defendant received an unexpectedly harsh sentence based on an AI risk score, but neither the judge nor defense counsel could understand how the score was calculated due to the model's complexity. This opacity prevents meaningful judicial review and appeals.

Accountability deficits compound interpretability problems. Developers blame biased input data, data providers cite algorithm design flaws, and judges disclaim responsibility by characterizing AI as merely an "aid" to decision-making. This diffusion of responsibility leaves biased outcomes unaddressed. When communities perceive AI systems as targeting them with unfair sentences without recourse, public trust erodes, reducing cooperation with law enforcement and increasing recidivism risks.

5. Case Studies

5.1. Successful AI-Assisted Bias Mitigation

[Jurisdiction Name] implemented an AI sentencing in 2019 following concerns about racial disparities. The system was developed through collaboration between legal experts, data scientists, and community representatives. Training data included 10 years of anonymized court records subjected to rigorous cleaning to correct misclassifications and remove duplicate entries. Under-represented groups were oversampled to ensure proportional representation in the dataset.

The AI system used a deep learning architecture with fairness-aware loss functions that explicitly penalized disparate outcomes across demographic groups. Judges received training on interpreting AI recommendations while maintaining judicial discretion. A 2022 evaluation found ethnic disparities in drug offense sentencing decreased from 20% to 5%, with a 20% reduction in sentence length variability. Judges reported the system highlighted overlooked factors, such as a young offender's prior rehabilitation success, leading to more appropriate community sentences instead of incarceration.

5.2. AI-Exacerbated Bias

[Another Jurisdiction] implemented an AI risk assessment tool in 2018 using five years of arrest and conviction data without adequate validation. The dataset reflected historical over-policing of low-income Hispanic neighborhoods, resulting in their over-representation. The system used a simple decision-tree algorithm that prioritized factors like prior arrests and residential instability without contextual adjustments.

The tool labeled African-American and Hispanic defendants as high-risk 40% more frequently than white defendants with similar criminal histories. Judges, lacking training on algorithmic limitations, heavily relied on these scores, resulting in 30% longer sentences for minority defendants in drug possession cases. Public outcry followed investigative reporting revealing these disparities, leading to legal challenges and eventual suspension of the tool pending revision.

5.3. Comparative Analysis

The case studies highlight critical success factors: high-quality, representative training data; fairness-aware algorithm design; transparency in operations; and stakeholder engagement including judge training. Failures resulted from using biased, unvalidated data; simplistic algorithms ignoring social context; lack of transparency; and uncritical judicial reliance on outputs. These demonstrate that AI's impact on sentencing equity depends on holistic implementation approaches that address data quality, algorithm design, and human-AI interaction.

6. Addressing the Bias Issues in AI-based Criminal Sentencing

6.1. Data Preprocessing and Cleaning Strategies

Comprehensive data collection across geographic regions and demographic groups prevents under-representation biases. Balancing techniques like oversampling under-represented groups or generating synthetic data ensures diverse case representation. Rigorous cleaning protocols correct mislabeled offenses through cross-verification with multiple sources (police reports, court transcripts, probation records).

Statistical methods like interquartile range analysis identify and address outliers-such as unusually harsh sentences-that might skew algorithmic learning. Ongoing data monitoring tracks representation changes over time, ensuring datasets remain representative as societal conditions and enforcement practices evolve. Anonymization techniques protect privacy while preserving demographic information necessary for fairness auditing.

6.2. Algorithmic Auditing and Improvement

Regular third-party audits assess both accuracy and fairness across demographic groups using metrics including disparate impact ratios, equalized odds, and statistical parity. Biases identified through auditing are addressed by rebalancing training data, adjusting variable weights, or modifying algorithm objectives to include fairness constraints.

Adopting fairness-aware machine learning techniques-such as adversarial debiasing or reweighted loss functions-explicitly minimizes disparate outcomes. Diverse development teams including legal experts, social scientists, and community representatives help identify hidden biases during design. Continuous improvement processes incorporate stakeholder feedback and emerging research on algorithmic fairness.

6.3. Establishing Legal and Ethical Frameworks

Legal regulations mandate transparency in AI systems, requiring disclosure of training data sources, algorithmic methodologies, and factor weights-building on GDPR's "right to explanation" for automated decisions. Liability frameworks clarify responsibilities for developers (design), data providers (quality), and judges (appropriate use).

Ethical guidelines prohibit using protected characteristics as inputs and restrict reliance on proxies for such characteristics. They emphasize proportionality and rehabilitation, requiring algorithms to consider program availability and participation potential. Independent oversight bodies with multidisciplinary expertise conduct regular audits, investigate complaints, and recommend improvements to AI systems and their implementation.

7. Conclusion

7.1. Summary of Key Findings

This research demonstrates AI's dual potential in criminal sentencing: reducing bias through data-driven objectivity (documented 8-point racial disparity reduction), enhanced consistency (decreased sentence variability), and comprehensive factor analysis; while risking bias amplification through flawed data (40% higher minority high-risk labeling), algorithm design issues (COMPAS racial disparities), and poor interpretability/accountability. Case studies confirm that AI's impact depends on data quality, fairness-aware design, transparency, and stakeholder engagement.

7.2. Implications for the Future

Future AI implementation in criminal justice requires prioritizing quality, representative data; designing algorithms for fairness alongside accuracy; ensuring transparency in operations; and developing legal/ethical frameworks that maintain human judgment as central. Judges need training to critically evaluate AI recommendations while leveraging their strengths. Responsible AI development must involve diverse stakeholders including affected communities to address contextual factors influencing fairness.

7.3. Directions for Further Research

Future work should develop robust supervision mechanisms like independent algorithmic auditors and blockchain verification systems. Research into AI's differential impact across crime types and demographic subgroups will enhance tool calibration. Developing user-friendly explainable AI interfaces will improve judicial understanding and oversight. Exploring broader ethical implications-including impacts on trust and due process-will inform comprehensive guidelines for equitable AI implementation in criminal justice.

Acknowledgments

I would like to express my sincere gratitude to all those who have contributed to the completion of this paper.

First and foremost, I am deeply indebted to my supervisor, [Supervisor's Name], whose invaluable guidance, insightful feedback, and unwavering support have been instrumental throughout the research and writing process. Their expertise in the fields of artificial intelligence and criminal justice has significantly shaped the direction and depth of this study.

I would also like to thank the faculty members and researchers at [Institution Name] for their constructive comments and suggestions during the various stages of this project. Their diverse perspectives have enriched the analysis and strengthened the arguments presented in this paper.

My appreciation extends to the legal professionals, technologists, and policymakers who generously shared their practical experiences and insights on the application of AI in criminal sentencing. Their firsthand knowledge has provided valuable context to the theoretical framework of this research.

I am grateful to the [Funding Body Name], if applicable, for providing financial support that enabled access to relevant data, resources, and academic conferences related to this study.

Additionally, I would like to acknowledge my peers and colleagues for their stimulating discussions and moral support. Their encouragement and collaborative spirit have been a source of motivation during challenging times.

Finally, I wish to express my heartfelt thanks to my family and friends for their patience, understanding, and unwavering encouragement throughout this journey. Their love and support have been crucial to the successful completion of this work.

References

[1] Alexander, M. (2010). The New Jim Crow: Mass Incarceration in the Age of Colorblindness. The New Press.

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

[3] American Bar Association. (2015). Report on the Uniformity of Sentencing. American Bar Association Criminal Justice Section.

[4] Berger, R. (2017). Reducing Racial Disparities in Sentencing through AI. Journal of Law and Technology, 45(2), 157-182.

[5] Dressler, J., & Michaels, S. (2009). Understanding Criminal Law (6th ed.). LexisNexis.

[6] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human Decisions and Machine Predictions. The Quarterly Journal of Economics, 133(1), 237-293. https://doi.org/10.1093/qje/qjx038