WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

The Paradox of Algorithmic Justice: Towards a Legally Defensible Framework for Algorithmic Fairness and Accountability in the Rule of Law

J ABDUL RAHMAN

abdulrahmanj965@gmail.com
Department of Computer Science and Engineering,
Sri Venkateswara College of Engineering, Chennai, India

A B S T R A C T :

The criminal justice system's increasing reliance on predictive algorithms, such as the COMPAS tool, is intended to enhance efficiency and objectivity. However, this approach has created a fundamental paradox of algorithmic fairness, where different, and often contradictory, definitions of fairness lead to disparate outcomes. The 2016 ProPublica investigation of the COMPAS algorithm exposed this by showing that while it met the standard of predictive parity, it failed the standard of equalized odds, demonstrating racially disparate error rates. This problem is not a simple technical bug but rather a reflection of historical and societal biases embedded in the training data, a phenomenon known as the "dirty data" problem. The legal system's traditional "fairness through unawareness" approach is ineffective in addressing this issue. A robust, multidisciplinary framework is needed, which includes mandated independent audits, legal standards informed by novel techniques like causal inference and peer-induced fairness, and a commitment to human oversight, as outlined in emerging regulations like the EU AI Act. The goal is to ensure that these tools support, rather than undermine, the rule of law and due process.

Keywords: Algorithmic Fairness, COMPAS, Predictive Policing, Algorithmic Bias, Algorithmic Accountability, Legal-Technical Divide, Due Process, EU AI Act, Equalized Odds, Predictive Parity

1. Introduction

1.1. The Shift to Algorithmic Governance

The criminal justice system is undergoing a significant transformation, increasingly adopting predictive algorithms to inform and optimize high-stakes decisions. This shift from subjective, experience-based human judgment to a more objective, data-driven approach is seen by proponents as a means to enhance efficiency, reduce costs, and improve the quality of justice.[1] Predictive policing algorithms, for instance, are designed to move law enforcement from a reactionary stance to a proactive one by forecasting where and when crimes are most likely to occur.[4] These systems analyze vast, sophisticated datasets that include decades of historical crime records, geographic information, and, in some cases, demographic details, to assign a "risk score" to particular locations or individuals.[4] Similarly, tools like the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) are widely used to predict recidivism risk for bail and sentencing decisions.[1] The fundamental premise is that modern AI systems, being "fancy calculators" that can flawlessly follow instructions, are uniquely suited for making predictions of human behaviour, a task that has historically been based on "crude, generalized, and non-tested assumptions".[1]

The notion that algorithms can remove human bias from the justice system is a seductive but dangerous oversimplification. This argument often overlooks the provenance of the training data itself. If an algorithm is trained on historical crime data that reflects pre-existing societal biases—such as the disproportionate policing and arrest rates in minority communities—it will learn to associate race or neighbourhood with higher risk.[4] The algorithm will then direct more police patrols to these same communities, leading to more arrests, which in turn feeds back into the algorithm as new data, reinforcing the initial, biased patterns.[7] This creates a "self-perpetuating cycle of prejudice" where the algorithm is not an impartial observer but a mirror reflecting and amplifying the discriminatory practices of the past.[4] This phenomenon has been termed "tech-washing," where racially biased policing methods are given a veneer of scientific objectivity simply because they are executed by a computer.[5]

1.2. Problem Statement:

The Unsolved Problem of Fair AI in JusticeWhile the problem of algorithmic bias is now widely acknowledged, the central challenge remains unsolved: developing a universally accepted and legally defensible framework for measuring, mitigating, and proving the absence of bias. This is because the very concept of "fairness" is not a singular, monolithic idea but a multi-disciplinary, and often contradictory, construct.[10] What a legal expert considers fair—due process, transparency, and non-discrimination based on protected characteristics—may differ from a computer scientist's quantitative definition.[11] This divergence is a key factor in the difficulty of developing a unified solution. The "black box" nature of many of these models further complicates this

issue, making it difficult for an individual to challenge a decision or for an auditor to assess its fairness.[8] This lack of transparency raises significant due process concerns, as a defendant may not be able to challenge how an algorithm reached a conclusion that affects their liberty.[8]

The fundamental challenge is not a technical failure but a conceptual and philosophical conflict. The debate over algorithmic fairness is a proxy for deeper societal disagreements about justice and equity. As demonstrated by a landmark study of the COMPAS algorithm, a model can be deemed "fair" by one definition but simultaneously "discriminatory" by another, and it is mathematically "impossible to make the algorithm fair in both ways" at the same time.[10] This critical insight transforms the issue from a solvable technical bug into a fundamental socio-legal dilemma. The question becomes, "What version of fairness do we, as a society, legally and morally prioritize?" This requires a framework that can not only measure bias but also provide a reasoned, legally robust justification for the trade-offs being made, with clear explanation and documentation.[11]

2. Case Study: The COMPAS Algorithm and the Fairness Paradox

2.1. The ProPublica Investigation

The 2016 ProPublica investigation into the COMPAS recidivism risk algorithm is a seminal moment in the public discourse on algorithmic bias.[6] The report brought the abstract concepts of "black box" bias and competing fairness definitions into sharp, real-world focus by analyzing data from more than 7,000 defendants in Broward County, Florida.[6] While the algorithm correctly predicted an offender's recidivism about 61% of the time, the investigation found that it made "mistakes in very different ways" for Black and white defendants.[15] Specifically, the analysis revealed that Black defendants who did not recidivate were nearly twice as likely to be misclassified as high-risk (45% of Black defendants vs. 23% of white defendants).[10] Conversely, white defendants who did recidivate were mistakenly labeled low-risk nearly twice as often as Black re-offenders (48% vs. 28%).[15] These findings laid bare the existence of a racial disparity in the model's error rates.

2.2. The Contradiction of Fairness Definitions

The debate that followed the ProPublica report perfectly illustrates the philosophical paradox of fairness. The company behind COMPAS, Northpointe, did not dispute the misclassification findings but rather defended its algorithm using a different definition of fairness.[10] This conflict highlights the impossibility of satisfying all fairness metrics simultaneously.

- **Equalized Odds (ProPublica's Critique):** This definition of fairness requires that an algorithm's error rates—specifically, the true positive rate (TPR) and the false positive rate (FPR)—are the same across protected groups.[10] ProPublica's analysis, as detailed above, proved that COMPAS failed this metric, as its misclassification errors were racially disparate.[15] A false positive, or an incorrect classification of a person as high-risk, occurred at a rate almost double for Black defendants.[15] A false negative, or an incorrect
- classification as low-risk, occurred at a rate almost double for white defendants.[15]
- **Predictive Parity (Northpointe's Defense):** This definition of fairness, also known as accuracy equity, asserts that an algorithm is fair if a given risk score means the same thing regardless of race.[10] Northpointe argued that for any given risk score (e.g., a score of 7 out of 10), the recidivism rate was "nearly identical" for Black and white defendants (61% of Black defendants re-offended vs. 60% of white defendants).[10] The company's defense was that their algorithm was equally accurate for all groups and therefore could not be biased.[10]

This central contradiction arises because while COMPAS satisfied a definition of predictive parity, it failed the test of equalized odds. This is not a technical flaw that can be easily debugged; it is a fundamental mathematical trade-off.[14] The inability to satisfy both definitions at once means that any algorithmic justice framework must make explicit choices and compromises.

2.3. Data Visualizations to Illustrate the COMPAS Paradox

The abstract concepts of these competing fairness definitions can be best understood through the specific data points from the ProPublica investigation. The following table and conceptual figure visually represent the numerical and distributional disparities.

Table 1: The COMPAS Fairness Paradox: Misclassification Rates by Race

This table demonstrates the failure of the COMPAS algorithm to satisfy the "Equalized Odds" metric. It quantifies the racial disparity in the model's errors, showing how it misclassified Black and white defendants in starkly different ways.

The data in Table 1 shows that Black defendants who did not re-offend were misclassified as high-risk at a rate nearly double that of their white counterparts. Conversely, white defendants who did re-offend were misclassified as low-risk at a rate nearly double that of their Black counterparts.[15]

Defendant Group	False Positive Rate (High Risk Misclassification)	False Negative Rate (Low Risk Misclassification)
Black Defendants	45	28
White Defendants	23	48

Table 1: The COMPAS Fairness Paradox: Misclassification Rates by Race

Figure 1: COMPAS Recidivism Score Distribution by Race

This conceptual histogram illustrates the racial disparity in the assignment of risk scores, which is the root cause of the fairness paradox. It visually demonstrates that the algorithm produced a different distribution of outcomes for each racial group, with Black defendants disproportionately assigned higher risk scores.

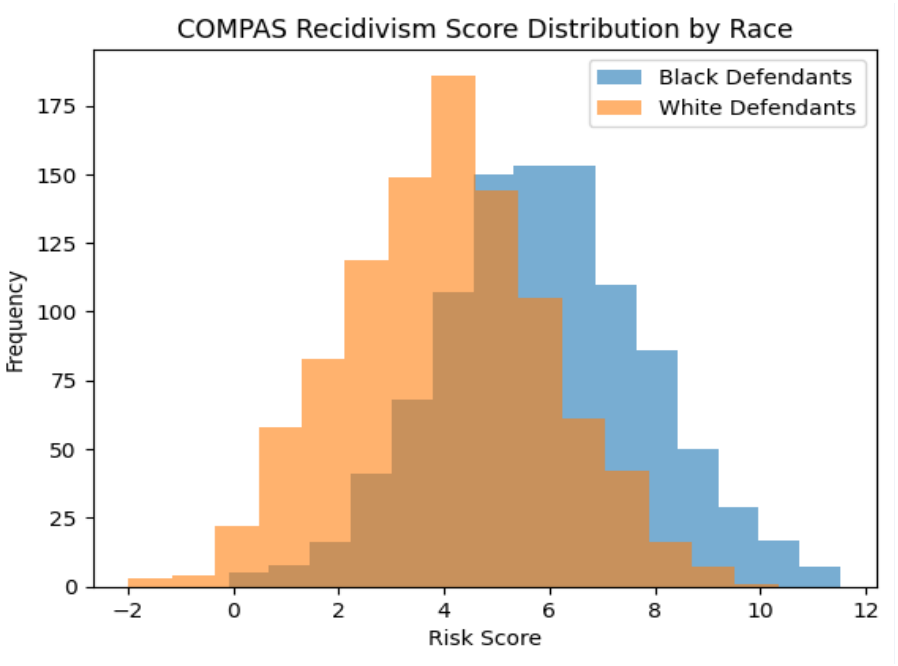

Figure 1: COMPAS Recidivism Score Distribution by Race

3. A Taxonomy of Algorithmic Fairness Metrics

The conflict over the COMPAS algorithm highlights the necessity of a nuanced understanding of fairness. In the data science community, fairness is not a single concept but a collection of metrics, often categorized as either group fairness or individual fairness.

3.1. Group Fairness Metrics

Group fairness metrics compare the performance of a model across predefined, legally protected groups.

Demographic Parity

Demographic parity is one of the simplest and most intuitive group fairness metrics. It is achieved when a model's prediction is statistically independent of membership in a sensitive group.[20] In a binary classification setting, such as a hiring or loan application system, this means that the proportion of applicants selected for a positive outcome should be equal across all groups, regardless of their true aptitude or risk profile.[18] This metric is often the easiest to implement and is useful for assessing "allocation harms," which occur when an AI system distributes opportunities or resources unequally.[20]

The simplicity of demographic parity is also its greatest weakness. Enforcing this metric can be deeply unjust because it fails to account for potentially different underlying "base rates" of outcomes between groups.[14] For example, in a health care context, if a model enforces demographic parity for a disease prediction, it might under-predict the disease in a group with a higher prevalence and over-predict it in a group with a lower prevalence to achieve a statistically equal outcome.[18] In a justice context, this could mean an algorithm would need to assign a lower risk score to defendants in a group with a higher base rate of recidivism to satisfy the metric, an action that may not be legally or ethically defensible. By only using predicted values and ignoring the true values, this metric can lead to a situation where a company hires qualified candidates from one group and careless ones from another at the same rate to satisfy the metric.[20]

Equalized Odds and Equal Opportunity

Equalized odds is a stricter group fairness metric that directly addresses the shortcomings of demographic parity. It requires that an algorithm's error rates—both the true positive rate (TPR) and the false positive rate (FPR)—are the same across all sensitive groups.[18] For a recidivism model, this would mean that the rate at which it correctly identifies re-offenders (TPR) and the rate at which it incorrectly identifies non-re-offenders (FPR) must be equal for all racial groups. Equal opportunity is a weaker variant of this metric, requiring only that the TPR is the same across groups, which is useful when false positives are considered less harmful than false negatives.[20] These metrics are considered stronger than demographic parity because they compare the model's performance against the ground truth, rather than simply its output.[18]

3.2. Individual Fairness Metrics

While group fairness metrics are valuable for assessing aggregate performance, they can obscure bias that affects individuals. Individual fairness metrics provide a complementary perspective.

Counterfactual Fairness

Counterfactual fairness is an individual-level metric that is concerned with whether two individuals who are identical in all respects except a sensitive attribute (e.g., race, gender) would receive the same model prediction.[21] For example, if a criminal defendant's risk score would be the same in a "counterfactual" world where their race was different, then the model is considered to be counterfactually fair. This provides a critical lens that aggregate group metrics can obscure, revealing systemic biases that may not be apparent when examining group averages.[22] For example, an admissions model could accept qualified candidates from a majority and a minority group at the same rate, satisfying demographic parity, but still reject the most qualified minority candidate while accepting a similarly qualified majority candidate.[22] This would be a clear violation of counterfactual fairness, which highlights the need for a multi-faceted approach to bias detection. A holistic view of bias requires using a combination of metrics, as no single metric is sufficient to ensure fairness.[22]

4. The Source of the Problem: Legal, Social, and Data-Driven Bias

4.1. The "Dirty Data" Problem

The root of algorithmic bias in the justice system is not the code itself but the historical and societal biases encoded in the training data.[4] This issue is a primary concern for predictive policing algorithms. These tools are trained on historical crime data that is often "derived from or influenced by corrupt, biased, and unlawful practices".[3] For example, studies on the predictive policing algorithm PredPol have shown that it led to the even greater over-policing of communities that were already patrolled at a high rate.[9] The NYPD's Patternizr, which is designed to not consider sensitive attributes like race or gender, has also been shown to compound implicit biases because the underlying data is already skewed.[9]

The existence of this "dirty data" problem is demonstrated by real-world policing practices. The NYPD's Stop-and-Frisk data provides a powerful empirical demonstration of this issue, illustrating how historical data can be deeply biased against minority communities and serve as a flawed foundation for algorithmic decision-making.

Figure 2: NYPD Stop-and-Frisk Racial Disparities (Conceptual)

This figure illustrates the racial disparities in the NYPD's stop-and-frisk program, which serves as a prime example of the type of historical data that can perpetuate bias in predictive policing algorithms.

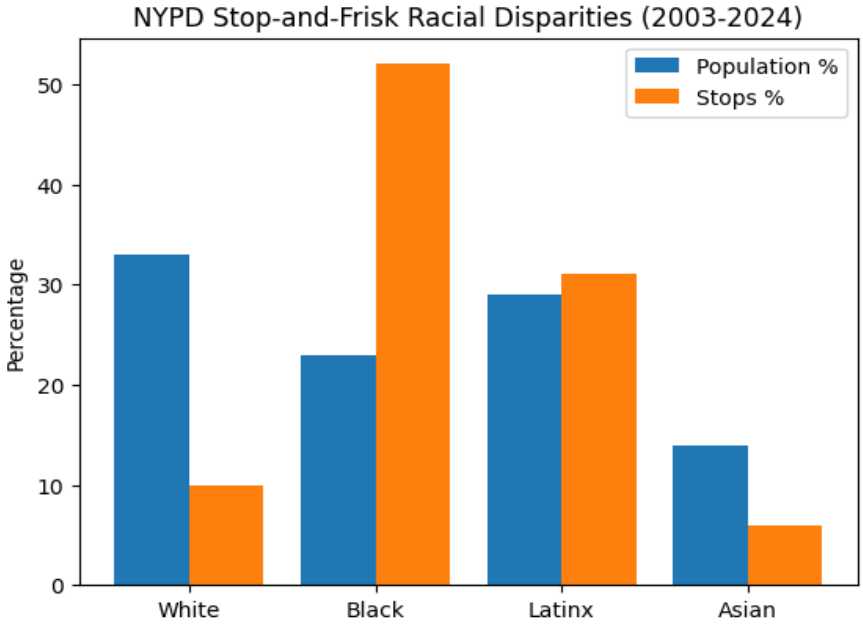

Figure 2: NYPD Stop-and-Frisk Racial Disparities (Conceptual)

The data presented in this figure shows that from 2003-2024, 90% of people stopped by the NYPD were people of color, despite representing a smaller percentage of the city's population.[23] Black New Yorkers, who make up 23% of the population, accounted for 52% of all stops, a rate nearly eight times greater than that of white people.[23] This kind of historical data, when fed into a predictive algorithm, would inevitably reinforce and automate these existing patterns of over-policing in minority communities, a phenomenon which is not the algorithm's fault but a reflection of a deeply biased system.[5]

4.2. The Legal-Technical Divide

A significant obstacle to a universally accepted framework is the conflict between technical debiasing methods and legal anti-discrimination principles.[24] Many algorithmic bias mitigation techniques, such as algorithmic recalibration, require the use of protected class variables or their proxies to correct for systemic bias.[24] However, U.S. anti-discrimination law has a "strong preference for decisions that are blind to them".[24] This legal approach, known as "fairness through unawareness" in the machine learning community, is widely considered "naive" and ineffective because proxies for protected attributes are almost always present in the data.[19]

The legal system's focus on input-blindness can create a dangerous legal "safe harbor" that permits biased algorithms to operate with impunity, as they can technically claim to be non-discriminatory.[24] The proposed HUD rule that would have granted a safe harbor to algorithms that did not use protected variables is a prime example of this legal-technical disconnect.[24] The law often requires a "causal connection" between a decision and a discriminatory outcome, but proving this is difficult with a "black box" algorithm.[24] The solution lies in using advanced techniques like causal inference to establish this connection, thus allowing for legally justifiable debiasing methods that may seem to run counter to traditional anti-discrimination law.[24]

5. Toward a Legally Defensible Framework for Auditable Justice

5.1. A Multi-disciplinary Approach to Accountability

A single technical fix or a narrow legal statute is insufficient to address the systemic problem of algorithmic bias. True accountability requires a holistic, multi-disciplinary framework that integrates legal, ethical, and technical principles.[25] This approach, as proposed by some researchers, can be grounded in international human rights law, which offers a framework for assessing harm across the "overall algorithmic life cycle" from design to deployment.[27] Accountability in this context means taking "corrective actions when algorithms cause harm or fail to operate as expected".[12]

5.2. Auditing and Transparency

The "black box" problem necessitates a fundamental shift in focus from scrutinizing proprietary source code to auditing the observable inputs and outputs of a model.[13] Mandatory, regular audits conducted by independent third-party organizations are a critical component of this framework.[12] These audits must verify model reliability, detect and mitigate biases, and ensure compliance with regulatory standards.[25] The *State v. Loomis* case in Wisconsin, where the court upheld the use of the COMPAS algorithm, highlighted the judiciary's "fundamental misunderstanding" of this issue, as the court focused on a request to view source code when the true problem was the biased data and its weights.[13] A robust auditing framework must be designed to circumvent these misunderstandings by providing clear, standardized, and legally defensible metrics for evaluation.[25]

5.3. Novel Approaches for Legally-Informed Fairness

The technical community is developing novel approaches to address the fairness paradox and provide a bridge between legal and technical definitions.

- **Causal Inference:** This technique moves beyond simple correlation to establish a causal link between an algorithm's decision process and a disproportionate outcome.[24] This aligns with legal requirements, offering a path to "reconcile technical feasibility and legal precedence".[24] By demonstrating that a particular model feature or training data pattern is a direct cause of discrimination, this approach provides the necessary evidence for a legal challenge.
- **Peer-Induced Fairness:** This novel framework aims to assess fairness at the individual level by comparing an individual's treatment to that of "similarly situated individuals" or peers.[31] It leverages causal inference and human notions of fairness based on social comparison.[31] This framework provides a powerful way to make algorithmic fairness intuitive and legally defensible by translating abstract statistical concepts into a human-centric comparison that a judge or a defendant can understand. The core contradiction of equalized odds and predictive parity in the COMPAS case stemmed from a fundamental misunderstanding of statistical measures that defy human intuition. The "peer-induced fairness" framework circumvents this by providing a simple, understandable metric: "was I treated like my peers?".[31] This approach could be key to addressing the core legal problem of due process and accountability, as it provides a clear, defensible standard for proving the absence of bias on a case-by-case basis, rather than relying on aggregate group statistics.

5.4. Review of Emerging Regulatory Frameworks

A global shift is underway from voluntary ethical guidelines to mandatory, legally-enforceable frameworks, signaling a consensus that self-regulation is insufficient.[27]

- **The Algorithmic Accountability Act (U.S.):** This proposed U.S. legislation would require companies using AI for "critical decisions" (e.g., employment, housing, legal services) to conduct impact assessments and report their findings to the Federal Trade Commission (FTC).[34] It aims to create transparency and accountability by documenting data inputs, performance, and mitigation strategies for negative impacts.[34]
- **The Eliminating Bias in Algorithmic Systems Act (U.S.):** Another proposed U.S. bill, this legislation would mandate that federal agencies using AI establish civil rights offices with a focus on bias and discrimination.[37] These offices would be required to submit regular reports to Congress on the state of AI technology and the steps taken to mitigate bias.[38]
- **The EU AI Act:** This legislation, which entered into force in August 2024, serves as a comprehensive global standard. It classifies certain AI systems—including predictive policing, evidence evaluation, and tools assisting judicial authorities—as "high-risk" due to their potential impact on fundamental rights.[2] The Act imposes strict requirements for data governance, risk management, documentation, and human oversight.[39]

Critically, it explicitly bans AI that makes decisions "solely" based on a person's profile and affirms that the "final decision-making must remain a human-driven activity".[2] The EU AI Act's rigorous classification and requirement for human-in-the-loop decisions are a crucial step toward protecting fundamental rights in high-stakes justice applications.

These frameworks, though nascent, are the most promising path toward a future where algorithmic accountability is a legal and technical reality, not just a theoretical ideal.

6. Conclusion

The proliferation of predictive algorithms in the justice system presents a profound challenge to the rule of law. The core problem is not a simple technical bug but a fundamental paradox: the competing and often mutually exclusive definitions of fairness. As the COMPAS case study demonstrates, an algorithm can be mathematically fair by one standard (predictive parity) while being discriminatory by another (equalized odds), exposing a deep conflict between different conceptions of justice. The root of this conflict is not malicious code, but the "dirty data" that reflects and amplifies existing societal biases. The legal system's traditional focus on "fairness through unawareness" and its inability to address the "black box" problem have only exacerbated this issue, creating a legal vacuum that allows biased algorithms to operate with impunity.

To address this multifaceted challenge, a multi-disciplinary framework for algorithmic accountability is required. This framework must move beyond simple bias detection and into the realm of legally defensible measurement, mitigation, and auditing. The following recommendations provide a path forward:

- **For Policymakers:** Legislation should be enacted that mandates independent, third-party audits for all high-risk algorithmic systems used in the justice sector. This legislation must move beyond "fairness through unawareness" and explicitly permit the use of race-conscious debiasing techniques when they are shown to be the most effective means of correcting systemic bias. The EU AI Act serves as a model for this approach by classifying high-risk systems and imposing strict requirements for human oversight.
- **For the Judiciary:** Judges and legal professionals should receive mandatory training on algorithmic risk assessment, including the nuances of different fairness metrics and the limitations of these tools. To uphold due process, they must scrutinize the inputs and outputs of a model and demand clear, human-understandable explanations for its decisions. The final judgment must always remain with the human judge, with the algorithm serving only as a supporting tool.
- **For Developers and Deployers:** The development community must embrace a "fairness-by-design" approach that incorporates ethical principles from the outset. This includes the continuous monitoring of models post-deployment to detect and correct for new biases that emerge over time. Novel techniques like causal inference and peer-induced fairness should be employed to design transparent and auditable systems. This approach provides a clearer, legally defensible standard for proving the absence of bias on a case-by-case basis, rather than relying on aggregate group statistics that can be easily manipulated or misunderstood.

The ultimate goal is to create a symbiotic relationship between law, ethics, and technology. Algorithms can be powerful tools for justice, but only if we embed them within a robust, legally defensible framework that acknowledges their limitations and prioritizes fundamental human rights and accountability above all.

REFERENCES

[1] Florida State University Law Review, "A framework for the efficient and ethical use of artificial intelligence in the criminal justice system," 2022. [Online]. Available: https://www.fsulawreview.com/wp-content/uploads/2022/08/ETHICAL-USE-OF-ARTIFICIAL-INTELLIGENCE.pdf. [Accessed: Aug. 29, 2025].

[2] eucrim, "Artificial intelligence and digitalisation of judicial cooperation," 2025. [Online]. Available: https://eucrim.eu/articles/artificial-intelligence-and-digitalisation-of-judicial-cooperation/. [Accessed: Aug. 29, 2025].

[3] Thomson Reuters Legal Solutions, "Predictive policing: Navigating the challenges," 2025. [Online]. Available: https://legal.thomsonreuters.com/blog/predictive-policing-navigating-the-challenges/. [Accessed: Aug. 29, 2025].

[4] Johns Hopkins Undergraduate Law Review, "Algorithmic justice or bias: Legal implications of predictive policing algorithms in criminal justice," Jan. 1, 2025. [Online]. Available: https://jhulr.org/2025/01/01/algorithmic-justice-or-bias-legal-implications-of-predictive-policing-algorithms-in-criminal-justice/. [Accessed: Aug. 29, 2025].

[5] Brennan Center for Justice, "Predictive policing explained," 2025. [Online]. Available: https://www.brennancenter.org/our-work/research-reports/predictive-policing-explained. [Accessed: Aug. 29, 2025].

[6] IRIS-BH, "Why does science need to talk more about ethics? A COMPAS case study and analysis of the presence of machine bias in automated decisions," 2025. [Online]. Available: https://irisbh.com.br/en/why-does-science-need-to-talk-more-about-ethics-a-compas-case-study-and-analysis-of-the-presence-of-machine-bias-in-automated-decisions/. [Accessed: Aug. 29, 2025].

[7] Reddit, "The ethics of policing algorithms," r/philosophy, 2025. [Online]. Available: https://www.reddit.com/r/philosophy/comments/z1wim3/the_ethics_of_policing_algorithms/. [Accessed: Aug. 29, 2025].

[8] University of Waterloo, "Use of AI in crime prediction algorithms: Ethical and legal implications," 2025. [Online]. Available: https://uwaterloo.ca/defence-security-foresight-group/sites/default/files/uploads/documents/khursheed_use-of-ai-crime-prediction-algorithms.pdf. [Accessed: Aug. 29, 2025].

[9] Yale Law School, "Algorithms in policing: An investigative packet," 2025. [Online]. Available: https://law.yale.edu/sites/default/files/area/center/mfia/document/infopack.pdf. [Accessed: Aug. 29, 2025].

[10] Columbia Law School, "Reprogramming fairness: Affirmative action in algorithmic criminal sentencing," 2025. [Online]. Available: https://hrlr.law.columbia.edu/hrlr-online/reprogramming-fairness-affirmative-action-in-algorithmic-criminal-sentencing/. [Accessed: Aug. 29, 2025].

[11] Berkeley Haas, "What does fairness mean for machine learning systems?," 2025. [Online]. Available: https://haas.berkeley.edu/wp-content/uploads/What-is-fairness_-EGAL2.pdf. [Accessed: Aug. 29, 2025].
[12] Rosenblum Law, "Algorithmic accountability in legal decisions," 2025. [Online]. Available: https://www.rosenblumlawlv.com/algorithmic-accountability/. [Accessed: Aug. 29, 2025].
[13] Harvard Law, "Algorithmic due process: Mistaken accountability and attribution in State v. Loomis," 2025. [Online]. Available: https://jolt.law.harvard.edu/digest/algorithmic-due-process-mistaken-accountability-and-attribution-in-state-v-loomis-1. [Accessed: Aug. 29, 2025].
[14] A. Downey, "Algorithmic fairness — Recidivism case study," 2025. [Online]. Available: https://allendowney.github.io/RecidivismCaseStudy/02_calibration.html. [Accessed: Aug. 29, 2025].
[15] ProPublica, "How we analyzed the COMPAS recidivism algorithm," 2025. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accessed: Aug. 29, 2025].
[16] Kaggle, "COMPAS recidivism racial bias dataset," 2025. [Online]. Available: https://www.kaggle.com/datasets/danofer/compass. [Accessed: Aug. 29, 2025].
[17] United States Courts, "False positives, false negatives, and false analyses: A rejoinder to Machine Bias," 2025. [Online]. Available: https://www.uscourts.gov/sites/default/files/80_2_6_0.pdf. [Accessed: Aug. 29, 2025].
[18] RSNA Journals, "Algorithmic fairness in machine learning," Aug. 25, 2023. [Online]. Available: https://pubs.rsna.org/page/ai/blog/2023/08/ryai_editorsblog082523. [Accessed: Aug. 29, 2025].
[19] O. Bastani, "Lecture 15: Fairness definitions," University of Pennsylvania, 2024. [Online]. Available: https://www.seas.upenn.edu/~obastani/cis7000/spring2024/docs/lecture15.pdf. [Accessed: Aug. 29, 2025].
[20] Fairlearn, "Common fairness metrics," 2025. [Online]. Available: https://fairlearn.org/main/user_guide/assessment/common_fairness_metrics.html. [Accessed: Aug. 29, 2025].
[21] Google Developers, "Machine learning glossary: Responsible AI," 2025. [Online]. Available: https://developers.google.com/machine-learning/glossary/responsible-ai. [Accessed: Aug. 29, 2025].
[22] Google Developers, "Counterfactual fairness," 2025. [Online]. Available: https://developers.google.com/machine-learning/crash-course/fairness/counterfactual-fairness. [Accessed: Aug. 29, 2025].
[23] NYCLU, "A closer look at stop-and-frisk in NYC," 2025. [Online]. Available: https://www.nyclu.org/data/closer-look-stop-and-frisk-nyc. [Accessed: Aug. 29, 2025].
[24] A. Xiang, "Algorithmic bias," Tennessee Law Review, 2025. [Online]. Available: https://ir.law.utk.edu/tennesseelawreview/vol88/iss3/5/. [Accessed: Aug. 29, 2025].
[25] A. Hall, "Legal requirements for annual AI model use audits," 2025. [Online]. Available: https://aaronhall.com/legal-requirements-for-ai-model-use-audits/. [Accessed: Aug. 29, 2025].
[26] arXiv, "Bridging research gaps between academic research and legal investigations of algorithmic discrimination," 2025. [Online]. Available: https://arxiv.org/html/2508.14954v1. [Accessed: Aug. 29, 2025].
[27] Z-Inspection, "International human rights law as a framework for algorithmic accountability," 2022. [Online]. Available: https://z-inspection.org/wp-content/uploads/2022/10/IHR-law-as-a-framework-for-algorithmic-accountability.pdf. [Accessed: Aug. 29, 2025].
[28] University of Essex, "Algorithmic accountability," 2025. [Online]. Available: https://www.essex.ac.uk/research-projects/human-rights-big-data-and-technology/algorithmic-accountability. [Accessed: Aug. 29, 2025].
[29] GOV.UK, "Auditing algorithms: The existing landscape, role of regulators and future outlook," 2025. [Online]. Available: https://assets.publishing.service.gov.uk/media/626910658fa8f523c1bc666c/DRCF_Algorithmic_audit.pdf. [Accessed: Aug. 29, 2025].
[30] D. Krause, "Addressing the challenges of auditing and testing for AI bias: A comparative analysis of regulatory frameworks," SSRN, 2025. [Online]. Available: https://papers.ssrn.com/sol3/Delivery.cfm/5050631.pdf?abstractid=5050631. [Accessed: Aug. 29, 2025].
[31] arXiv, "Peer-induced fairness: A causal approach for algorithmic fairness auditing," 2024. [Online]. Available: https://arxiv.org/html/2408.02558v2. [Accessed: Aug. 29, 2025].
[32] University of Pennsylvania, "Peer-induced fairness in games," 2025. [Online]. Available: https://repository.upenn.edu/bitstreams/8cbafd1e-a488-4b9c-af93-5a97150c8e92/download. [Accessed: Aug. 29, 2025].
[33] BSA, "Statement on the Algorithmic Accountability Act of 2023," 2023. [Online]. Available: https://www.bsa.org/news-events/news/bsa-statement-on-the-algorithmic-accountability-act-of-2023. [Accessed: Aug. 29, 2025].
[34] UNC Law, "The Algorithmic Accountability Act and the future of algorithmic regulation," 2025. [Online]. Available: https://journals.law.unc.edu/ncjolt/blogs/the-algorithmic-accountability-act-and-the-future-of-algorithmic-regulation/. [Accessed: Aug. 29, 2025].
[35] Senate of the United States, "Algorithmic Accountability Act of 2023 Summary," 2023. [Online]. Available: https://www.wyden.senate.gov/imo/media/doc/algorithmic_accountability_act_of_2023_summary.pdf. [Accessed: Aug. 29, 2025].
[36] American Bar Association, "The Algorithmic Accountability Act," 2022. [Online]. Available: https://www.americanbar.org/content/dam/aba/publications/antitrust/magazine/2022/august/algorithmic-accountability-act.pdf. [Accessed: Aug. 29, 2025].
[37] Congress.gov, "S.3478 - Eliminating Bias in Algorithmic Systems Act of 2023," 2023. [Online]. Available: https://www.congress.gov/bill/118th-congress/senate-bill/3478. [Accessed: Aug. 29, 2025].
[38] Congress.gov, "Text - S.3478 - 118th Congress (2023-2024): Eliminating Bias in Algorithmic Systems Act of 2023," 2023. [Online]. Available: https://www.congress.gov/bill/118th-congress/senate-bill/3478/text. [Accessed: Aug. 29, 2025].
[39] GNET, "The EU's AI Act: Implications on justice and counter-terrorism," Mar. 10, 2025. [Online]. Available: https://gnet-research.org/2025/03/10/the-eus-ai-act-implications-on-justice-and-counter-terrorism/. [Accessed: Aug. 29, 2025].