Machine bias: new experiments with COMPAS data*

Ana Farič†
Faculty of Education
University of Ljubljana
Slovenia
af27987@student.uni-lj.si

Ivan Bratko
Faculty of Computer and Information Science
University of Ljubljana
Slovenia
bratko@fri.uni-lj.si

Abstract

This paper revisits the debate on machine bias through an analysis of the COMPAS recidivism prediction system. While some studies claim COMPAS is racially biased and others argue the opposite, our replication and extension of prior work show that across diverse methods accuracy consistently converges at around 66-67%. Moreover, error distributions follow a stable pattern: higher false positive rates for black defendants and higher false negative rates for white defendants. We argue that this convergence reflects inherent difficulty of this prediction problem and probably yet unexplained asymmetries in this domain. Our findings suggest that debates on fairness should move beyond model choice to address systemic disparities that shape observed outcomes.

Keywords

artificial intelligence, machine bias, fairness, COMPAS system

1 Introduction

Calls for unbiased AI systems are increasingly more common in regulation debates. For example, in September 2024, one of the GPAI (*Global Partnership on Artificial Intelligence*) working groups released a report [13] recommending that AI system providers be held liable for discriminatory impacts and required to compensate individuals harmed by algorithmic bias. Although the group attempted to clarify and better define the notion of bias in the revised report [14], released in November 2024, it ultimately offered no practical metrics or other criteria to determine with confidence whether a system is biased or not. This highlights a broader challenge: while the demand for unbiased AI systems is growing, even well-intentioned policymakers struggle to translate abstract concepts of fairness into actionable, measurable criteria. The COMPAS recidivism prediction system exemplifies these definitional difficulties. Some studies claim COMPAS is racially biased, while others disagree, depending on which fairness metric is applied. Beyond the technical debate, this inconsistency raises a deeper question: to what extent do observed disparities reflect the various models versus the underlying data and social context?

*Article Title Footnote needs to be captured as Title Note
†Author Footnote to be captured as Author Note

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia

http://doi.org/10.70314/is.2025.cogni.13

2 Understanding the concept of machine bias

In computer science, dozens of fairness metrics and bias definitions exist, often in contradiction with one another [16, 17, 18, 19]. Philosophers, legal scholars, and social scientists have long debated the meaning of bias, and computer scientists face the additional challenge of operationalizing these abstract concepts into measurable criteria [12]. Despite numerous attempts to resolve this ambiguity, no consensus has emerged. Even mathematical definitions, while precise, often lack concrete examples that would make them applicable in real-world decision-making contexts.

Bias in machine learning (ML) is a multifaceted concept, encompassing both technical and social dimensions. Researchers identify three broad types of bias:
1. Inductive/learning bias: in supervised learning, an algorithm seeks a function that predicts outcomes from data. Many functions may fit the training data, but most fail to generalize well. Preferential bias is needed to select certain functions over others, guiding learning toward useful generalizations. As such, bias is a necessary component that enables learning [16].
2. Historical bias: reflects real-world prejudices embedded in the data. As such, even perfectly measured data may produce biased outcomes if the underlying reality is discriminatory [1, 16, 19].
3. Biases that arise during data generation: specification, measurement/observation, sampling/population, annotator bias etc. [5, 16, 24].
In practice, bias is most often discussed in terms of its social consequences, such as when models classify individuals differently based on protected attributes like race or gender [16].

Computer scientists have formalized bias through various fairness metrics, including:
1. Demographic parity: equal positive prediction rates across groups [16];
2. Equalized odds: equal false positive (FPR) and false negative rates (FNR) across groups [15];

3. Predictive parity: equal prediction accuracy across groups [23].
These metrics are mutually incompatible with the exception of certain very trivial conditions [17].
For a more comprehensive survey of existing bias definitions and their limitations, we refer the reader to our previous work [9].

3 COMPAS

A good example that demonstrates the problem of a lack of a unified definition is COMPAS (*Correctional Offender Management Profiling for Alternative Sanctions*). It is a model still used by US courts to assess the likelihood that a defendant will reoffend within two years of the evaluation date if released [2, 20]. The assessment is based on 137 attributes about the defendant, including personal information and criminal history. Race is not considered in the evaluation [8]. The model assists judges in making decisions about bail and sentencing, particularly in determining whether defendants awaiting trial are too dangerous to be released [6].

The model assigns defendants a score between 1 and 10, indicating how likely they are to reoffend. These scores have a significant impact on the lives of the defendants. Those rated as medium- or high-risk (scores 5-10) are more often held in detention until trial, while low-risk defendants (scores 1-4) are more frequently released [2, 6].

3.1 Previous research on COMPAS

In 2016, ProPublica [2] sparked an intense debate by claiming that COMPAS was biased against black defendants. Over the following years, researchers reached contradictory conclusions, highlighting the difficulty of assessing bias in this system. Northpointe, the developer of COMPAS, rejected ProPublica's claims, arguing that ProPublica's analysis was methodologically flawed, and that ProPublica should have used standard fairness measures such as AUC-ROC, under which COMPAS showed no racial bias [7]. Similarly, Flores et al. [10] argued that there is no significant difference in predictive accuracy between white and black defendants. In AI literature, the negative assessments of COMPAS prevail.

Subsequent studies further complicated the debate. Dressel and Farid [8] showed that COMPAS performs no better than laypeople in predicting recidivism, and that a simple linear classifier with only two or seven attributes produces accuracy results comparable to COMPAS's 137-attribute system. Rudin [22] reached a similar conclusion with a three-rule interpretable model based on just two attributes. These findings questioned the added value of complex risk assessment tools in this domain.

Other research emphasizes inherent trade-offs in fairness metrics. Corbett-Davies et al. [6] and Zafar et al. [27] all highlight the impossibility of simultaneously satisfying competing fairness definitions, given differing base rates of recidivism across racial groups and different fairness metrics. Conceptual and formal tensions such as these, help explain why analyses of COMPAS produce conflicting assessments of the same system [9].

4 Our analysis

We noticed an intriguing pattern in the previously mentioned studies: models of varying complexity produce comparable accuracy, and FNR and FPR. While these results have been reported independently, they have not been systematically compared within the same analytical framework, nor have their broader implications been thoroughly examined. In this paper, we aim to replicate prior findings to confirm their robustness, and to extend the discussion by investigating why such convergence occurs across different methods.

4.1 Method

We used the publicly available COMPAS dataset released by ProPublica [2] on GitHub (https://github.com/propublica/compas-analysis). To ensure comparability with previous studies, we selected the same version of the dataset as used by [8]. The dataset contains 53 attributes, including demographic information, criminal history, and COMPAS risk scores, including protected attributes such as race and sex, for 7214 defendants from Broward County, Florida. Following the previous researchers, we filtered the dataset to include only black and white defendants, resulting in final 6150 individuals.

We trained the following models using the Orange data mining platform, applying an 80%/20% training/testing split that was repeated 10 times to compute true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), both overall and separately for black and white defendants:
1. Logistic regression: simple linear classifier, trained with either 6 (sex, age, prior crimes, crime degree, number of juvenile misdemeanors and felonies) or 2 attributes (age and priors). We excluded the crime description attribute (used by [8]) because it contains over 400 different values, which they reduced to 63 for human judgement purposes. As we could not reproduce this exact transformation, we omitted it to test whether the remaining attributes alone suffice.
2. Decision tree: was constructed to approximate Rudin's [22] rule-based model. Two attributes (age category and priors) were used and its depth was limited to 5.

Models were evaluated using:
1. Accuracy: proportion of correct predictions on test set.
2. FPR: proportion of non-recidivists incorrectly predicted to reoffend.
3. FNR: proportion of recidivists incorrectly predicted not to reoffend.
For each model, TP, TN, FP, and FN were first recorded for each of the 10 repetitions. These counts were then pooled

across repetitions, and accuracy, FPR and FNR were calculated from the pooled counts for each race group. Finally, metrics were averaged across all repetitions to produce values reported in tables 1 and 2.

The results from our models were directly compared to reported metrics from [2, 8, 22], allowing us to compare predictive performance across models.

4.2 Results

The results from previous researchers are summarized in table 1. Our results are summarized in table 2.

Across all methods (table 2), overall accuracy converged around 66-67%, consistent with the performance reported by ProPublica [2] and Dressel and Farid [8]. While our exact error rates differ somewhat from those reported previously, the same pattern in error distribution was observed; black defendants exhibited higher FPR, whereas white defendants exhibited higher FNR. Finally, compared to [8], who incorporated a reduced version of the *crime description* attribute, our results suggest that excluding this feature does not substantially change performance.

Table 1: Columns A-E summarize predictive performance across different models and conditions from previous researchers. Column A reports human judgements without access to information about race; B reports human judgements with race, C shows COMPAS predictions as reported by ProPublica, D and E show logistic regression (LR) models trained on 7 or 2 attributes respectively. Accuracy (CA), FPR and FNR are reported overall and separately for two races.

	A: human (no race)	B: human (race)	C: COMPAS	D: LR-7	E: LR-2
CA (overall)	67.0%	66.5%	65.2%	66.6%	66.8%
CA (black)	68.2%	66.2%	64.9%	66.7%	66.7%
CA (white)	67.6%	67.6%	65.7%	66.0%	66.4%
FPR (black)	37.1%	40.0%	40.4%	42.9%	45.6%
FPR (white)	27.2%	26.2%	25.4%	25.3%	25.3%
FNR (black)	29.2%	30.1%	30.9%	24.2%	21.6%
FNR (white)	40.3%	42.1%	47.9%	47.3%	46.1%

Table 2: Columns A-D summarize predictive performance of our models. A shows LR trained on 6 attributes, excluding race, B shows LR trained on the same 6 attributes with race included; C shows LR trained on 2 attributes, D shows a decision tree (DT) trained on 2 attributes. CA, FPR and FNR are reported overall and separately for black and white defendants.

	A: LR-6 (no race)	B: LR-6 (race)	C: LR-2	D: DT-2
CA (overall)	67.2%	67.1%	66.5%	66.8%
CA (black)	66.9%	67.2%	66.7%	66.8%
CA (white)	67.1%	66.4%	66.1%	67.7%
FPR (black)	29.0%	31.1%	31.1%	35.2%
FPR (white)	15.1%	15.2%	16.5%	20.1%
FNR (black)	36.7%	34.5%	35.4%	31.3%
FNR (white)	61.7%	61.4%	60.6%	51.0%

Figures 1-3 present an example of a decision tree trained on two attributes (age category: 1. < 25, 2. 25-45, 3. > 45) and number of prior offences, with the tree depth limited to 5. The tree splits defendants into subgroups, with the leaves representing predicted recidivism risk (0=predicted not to reoffend, 1=predicted to reoffend) and the proportion of majority class.

To improve readability, the tree is divided into three figures: figure 1 shows the root node and the initial split by the number of priors, figure 2 shows the left subtree (defendants with less or equal 2 priors), and figure 3 shows the right subtree (defendants with > 2 priors). Among these, defendants older than 45 are further divided by priors. The utmost right leaf in figure 3 predicts that defendants with more than 20 priors will reoffend (1), with probability 76.9%.

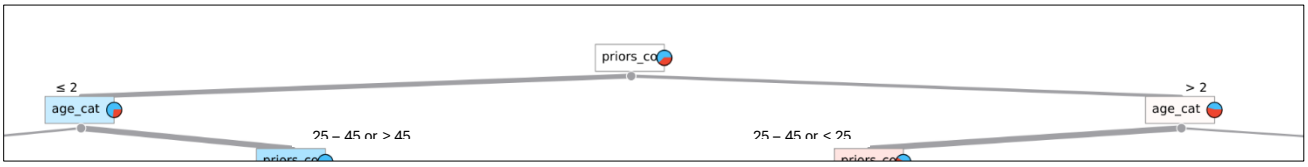

Figure 1: Root node of the decision tree and the initial split into left and right subtrees.

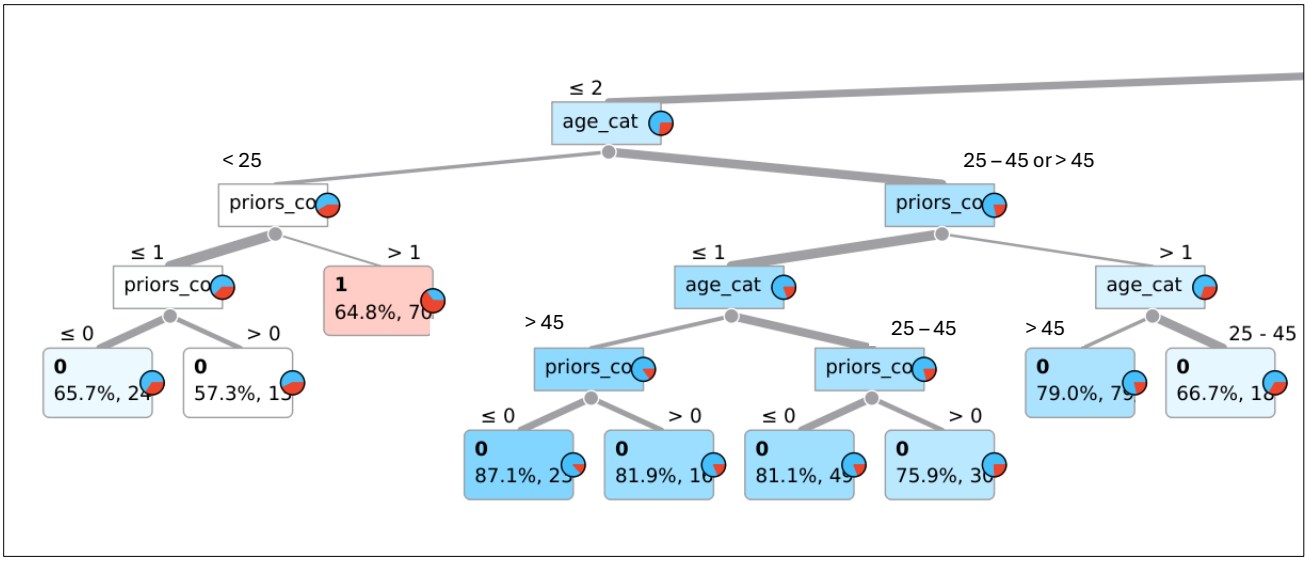

Figure 2: Left subtree of the decision tree

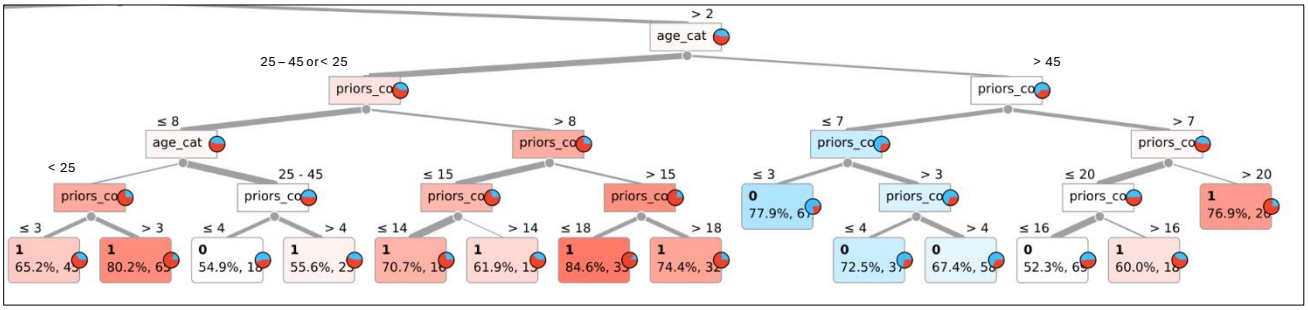

Figure 3: Right subtree of the decision tree

To further illustrate how the decision tree makes predictions, figures 4 and 5 show the distribution of defendants by race (the first two columns in both figures correspond to black defendants, the right ones to white defendants) within two example leaves. The bar charts make it clear that, although the prediction is the same within each leaf (1 = will reoffend, shown in red; the blue color indicates the number of defendants who did not reoffend), the underlying racial composition of these subgroups can vary substantially. Additionally, the figures illustrate how estimated prediction errors and class balance differ across leaves.

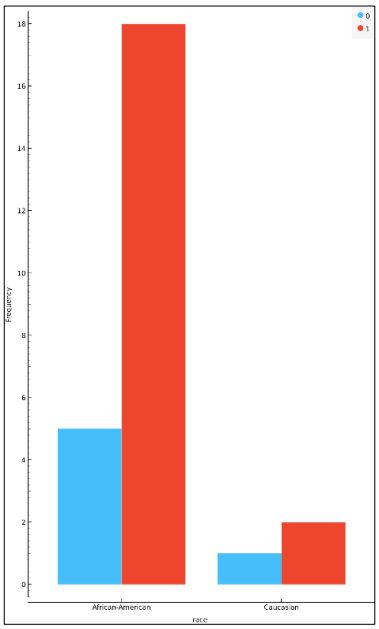

Figure 4: Distribution of defendants by race in the leaf (right-most leaf in figure 3) corresponding to the path: > 2 priors --> age > 45 --> > 20 priors. All defendants in this subgroup are predicted to reoffend, with 76.9% probability.

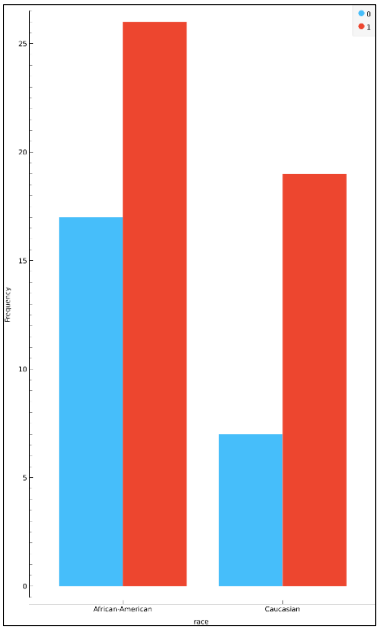

Figure 5: Distribution of defendants by race in the leaf (left-most leaf in figure 3) corresponding to the path: > 2 priors --> age < 25 --> 3 priors. All defendants in this subgroup are predicted to reoffend with probability 65.2%.

4.3 Discussion

Our findings confirm and extend those of [2, 8, 22]. Across models of varying nature and complexity (black box, logistic regression, interpretable rule-based, and even human judgement) predictive accuracy consistently hovers around 66-67%. Moreover, we replicated the characteristic error distribution pattern; higher FPR for black defendants, and higher FNR for white defendants. We extend the mentioned prior research by demonstrating this convergence using decision trees as an approximation of Rudin's [22] interpretable rules.

While [22] emphasizes the use of inherently interpretable models, and [8] question the overall utility of algorithmic recidivism prediction, our work shifts the discussion toward the underlying reasons why all these methods yield similar results, in particular similar error patterns.

The convergence of predictions across methods suggests that the limitations may lie less in model choice and more in the data and domain itself, which prior analyses often overlook.

Beyond dataset quality, structural factors such as racial disparities in arrests and sentencing likely drive the consistent error patterns observed across all models. As [11] reports, the lifetime likelihood of imprisonment for black men was one in three for those born in 1981, and one in five for those born in 2001. A report from 2018 [21] emphasizes that the imprisonment rate for black adults is 5.9 the rate for white adults – and even higher in some states. These disparities exist for both least and more serious offences; 56% of people imprisoned nationwide for a drug offence are black or Latino, and 48% of people serving life sentences are black. Another report [25] emphasizes that 56.4% of those serving life without parole sentences are black.

Additionally, Williamsons' framework [26] emphasizes that the higher crime rates observed among black individuals are not indicative of inherent crime tendencies, but rather reflect systemic economic disparities which are often the result of historical and ongoing policies that have marginalized black communities, limiting their access to resources and opportunities. Therefore, the convergence of predictive models like COMPAS with other simple ML models and even lay people judgements may not solely be a technical issue but also a reflection of deeper societal inequalities.

While our study confirms agreement across models and highlights the importance of structural factors, several avenues for further research remain. At a methodological level, further work could explore different versions of the ProPublica dataset, test additional feature combinations, and evaluate a wider range of ML models to assess the robustness of these patterns. At a broader level, additional research should examine the underlying systemic factors that drive disparities in recidivism predictions, thus contextualizing algorithmic predictions within real-world social dynamics and inform policy discussions on the responsible use of predictive models in the justice system.

5 Conclusion

Our study revisits the COMPAS recidivism prediction debate by replicating and extending previous findings and discussion why different methods (ranging from black-boxes to simple linear predictors, interpretable rule-based models, and human judgements) consistently converge on similar predictive performance and error patterns. Across all methods, accuracy hovered around 66-67%, with characteristic error distributions showing higher FPR for black defendants and higher FNR for white defendants.

While prior research has documented this convergence, its proper interpretation and broader implications have received less attention. We emphasize that COMPAS may be wrongfully vilified; while skepticism regarding algorithmic risk assessment is warranted, using ML systems to inform decisions should not be dismissed outright, as they hold potential to support informed decision-making if used responsibly.

More importantly, we argue that the debate should shift toward understanding why such convergence occurs. Our findings suggest that it reflects domain-specific and structural factors, including disparities in arrest, sentencing, and systemic socio-economic inequalities that induce observed recidivism rates. By examining these elements, alongside limitations in commonly used datasets, we can better contextualize predictive performance and the persistence of racial disparities.

References

[1] Alelyani, S. 2021. Detection and Evaluation of Machine Bias. *Applied Sciences* 11, 4. DOI:https://doi.org/10.3390/app11146271.

[2] Angwin, J., Kirchner, L., Larson, J. & Mattu, S. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica.*

[3] Barenstein, M. 2019. ProPublica's COMPAS Data Revisited. *ArXiv:1906.04711.* DOI: https://doi.org/10.48550/arXiv.1906.04711.

[4] Beck, A. J. 2021. *Race and Ethnicity of Violent Crime Offenders and Arrestees, 2018.* U.S. Department of Justice, Statistical Brief.

[5] Chakraborty, J., Majumder, S. & Menzies, T. 2021. Bias in Machine Learning Software: Why? How? What to do? *ESEC/SIGSOFT FSE.* DOI:https://doi.org/10.1145/3468264.3468537.

[6] Corbett-Davies, S., Pierson, E., Feller, A. & Goel, S. 2016. A computer program used for bail and sentencing decisions was labeled against blacks. It's actually not that clear. *The Washington Post.*

[7] Dieterich, W., Mendoza, S. & Brennan, T. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.

[8] Dressel, J. & Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4, 1. DOI:https://doi.org/10.1126/sciadv.aao5580.

[9] Farič, A. & Bratko, I. 2024. Machine Bias: A Survey of Issues. *Informatica* 48, 2. DOI: https://doi.org/10.31449/inf.v48i2.5971.

[10] Flores, A. W., Lowenkamp, C. T. & Bechtel, K. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder To "Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Federal Probation* 80, 2.

[11] Ghandnoosh, N. 7.12.2023. *One in Five: Racial Disparity in Imprisonment – Causes and Remedies.* The Sentencing Project. https://www.sentencingproject.org/reports/one-in-five-racial-disparity-in-imprisonment-causes-and-remedies/.

[12] Goel, N., Yaghini, M. & Faltings, B. 2018. Non-Discriminatory Machine Learning through Convex Fairness Criteria. *The 23nd AAAI Conference on Artificial Intelligence* 32, 1.

[13] GPAI 2024. *Towards Substantive Equality in Artificial Intelligence: Transformative AI Policy for Gender Equality and Diversity.* Report, September 2024, Global Partnership on AI.

[14] GPAI 2024. *Towards Substantive Equality in Artificial Intelligence: Transformative AI Policy for Gender Equality and Diversity.* Report, November 2024, Global Partnership on AI.

[15] Hardt, M., Price, E. & Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *ArXiv:1610.02413.* DOI:https://doi.org/10.48550/arXiv.1610.02413.

[16] Hellstrom, T., Dignum, V. & Bensch, S. 2020. Bias in Machine Learning – What is it Good for? *ArXiv:2004.006686.* DOI:https://doi.org/10.48550/arXiv.2004.00686.

[17] Kleinberg, J., Mullainathan, S. & Raghavan, M. 2021. Inherent Trade-Offs in the Fair Determination of Risk Scores. *ArXiv:1609.05807.* DOI:https://doi.org/10.48550/arXiv.1609.05807.

[18] Mehrabi, A., Morstatter, F., Saxena, N., Lerman, K. & Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys.* DOI:https://doi.org/10.1145/3457607.

[19] Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M. E., … Staab, S. 2020. Bias in data-driven artificial intelligence systems – An introductory survey. *WIREs Data Mining and Knowledge Discovery* 10, 3. DOI:https://doi.org/10.1002/widm.1356.

[20] Porębski, A. 2023. Machine learning and law. *Research Handbook on Law and Technology.* Cheltenham, UK: Edward Elgar Publishing.

[21] *Report to the United Nations on Racial Disparities in the U.S. Criminal Justice System.* 19.4.2018. The Sentencing Project. https://www.sentencingproject.org/reports/report-to-the-united-nations-on-racial-disparities-in-the-u-s-criminal-justice-system/.

[22] Rudin, C. 2018. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *ArXiv:1811.10154.* DOI:https://doi.org/10.48550/arXiv.1811.10154.

[23] Saravanakumar, K. K. 2021. The Impossibility Theorem of Machine Fairness: A Causal machine learning algorithms interaction. *ArXiv:2007.06024.* DOI:https://doi.org/10.48550/arXiv.2007.06024.

[24] Sun, O., Nasraoui, O. & Shafto, P. 2020. Evolution and impact of bias in human and machine learning algorithms interaction. *PLOS ONE* 15, 8. DOI:https://doi.org/10.1371/journal.pone.0235502.

[25] Walsh, A. 15.8.2016. *The criminal justice system is riddled with racial disparities.* Prison Policy Initiative. https://www.prisonpolicy.org/blog/2016/08/15/cjrace/.

[26] Williamson Kramer, C. 13.2.2024. *Systemic Racism in Crime: Do Blacks Commit More Crimes Than Whites?* Liberty Matters. https://oll.libertyfund.org/publications/liberty-matters/2024-02-13-systemic-racism-in-crime-do-blacks-commit-more-crimes-than-whites.

[27] Zafar, M. B., Valera, I., Gomez Rogrigues, M. & Gummadi, K. P. 2017. Fairness constraints: Mechanisms for fair classification. *ArXiv:1507.05259.* DOI:https://doi.org/10.48550/arXiv.1507.05259.