Cultural Bias in Machine Learning Systems: A Philosophical and Empirical Study of Algorithmic Knowledge Production

Nabulongo Ali., Peter Both Goah Wiech., Katwesigye Collins., Prof. Specioza Asiimwe

Kampala International University, Uganda

DOI: https://doi.org/10.47772/IJRISS.2026.100300368

Received: 12 March 2026; Accepted: 17 March 2026; Published: 09 April 2026

ABSTRACT

Machine learning systems are increasingly functioning as epistemic infrastructures in high-stakes domains such as criminal justice, healthcare, finance, and employment. Despite this, their outputs are frequently treated as objective and neutral forms of knowledge. This study advances a synthesis of empirical and philosophical inquiry into cultural bias in machine learning, arguing that algorithms operate as sociotechnical agents embedded within historically situated structures of power and representation. Using the COMPAS Recidivism dataset (N = 7,214), a quantitative experimental design was employed to examine predictive disparities across protected attributes, specifically race and sex. Logistic Regression and Random Forest models were implemented within a controlled preprocessing pipeline and evaluated using standard performance metrics (accuracy, precision, recall, and F1-score), alongside subgroup fairness measures including false positive rates (FPR), false negative rates (FNR), and disparate impact ratios. To ensure robustness, subgroup disparities were further assessed using statistical significance testing. While overall model performance was moderate in aggregate metrics, subgroup analysis revealed consistent and structured disparities: African-American defendants exhibited elevated false positive rates, whereas females and underrepresented racial groups experienced disproportionately high false negative rates. These patterns persisted across model architectures, indicating that bias is structurally embedded in the data rather than solely a function of model design. However, extreme subgroup values should be interpreted with caution due to potential sample size imbalances within certain demographic categories. The findings challenge the assumption of epistemic neutrality in algorithmic systems, demonstrating that machine learning models participate in the cultural production of knowledge by reproducing historically grounded classifications and power asymmetries. The study argues that algorithmic outputs should be evaluated not only in terms of predictive performance but also through fairness-aware and context-sensitive frameworks that account for their broader ethical and epistemological implications.

Keywords: Algorithmic Bias; Machine Learning Fairness; Cultural Bias; Disparate Impact; Recidivism Prediction; COMPAS Dataset; Ethical AI Governance; Epistemic Injustice

INTRODUCTION

Machine learning (ML) systems are now an important part of the decision-making systems used in finance, health care, education, employment, and criminal justice. People are relying on these structures more and more to give them information that can be used to make decisions, like whether someone is likely to pay back a loan or whether they are a good candidate for a job or a police officer. However, the notion that ML systems function as neutral or exclusively technical inference mechanisms has been substantially contested in recent academic discourse. Instead of being clear tools for knowledge, machine learning (ML) algorithms often encode and repeat patterns that reflect cultural and structural bias found in their training data. This leads to results that unfairly hurt some demographic groups more than others (Mehrabi et al., 2021; Barocas, Hardt & Narayanan, 2023).

Recent research indicates that algorithmic bias is not simply a technical anomaly but a systemic issue resulting from the interplay of data, model architecture, and social context. Bias in machine learning (ML) refers to systematic and repeatable errors that lead to unfair outcomes for certain social groups, often mirroring historical inequalities inherent in the data generation process (Mehrabi et al., 2021). These distortions may arise from representation bias, measurement bias, aggregation bias, or historical bias, all of which are mediated by

sociotechnical factors rather than being solely computational issues (Suresh & Guttag, 2021). Thus, training ML models on extensive datasets is never epistemically neutral; datasets embody culturally specific classifications, normative assumptions, and power imbalances. In this regard, ML systems contribute to the reproduction of social knowledge rather than merely identifying neutral patterns (Birhane, 2021).

The growing use of ML systems in important areas has made people more worried about fairness and accountability. Empirical research indicates that predictive systems can attain elevated overall accuracy while concurrently generating substantial disparities among demographic groups when assessed through subgroup-specific error rates (Friedler et al., 2021; Barocas et al., 2023). Consequently, fairness research prioritizes metrics such as false positive rate parity, equal opportunity, and disparate impact to uncover concealed inequities obscured by aggregate performance scores. In areas like credit scoring, facial recognition, and risk assessment, studies have shown that predictive error distributions differ across race, gender, or socioeconomic groups (Mehrabi et al., 2021; Birhane, 2021). These results indicate that algorithmic outputs cannot be regarded merely as objective predictions; instead, they must be contextualized within the overarching sociocultural conditions that influence data generation and model training.

From a philosophical perspective, this empirical evidence prompts essential epistemological inquiries. Traditional analytic epistemology frequently defines knowledge as justified true belief, highlighting the importance of reliability and objectivity in acquisition methods. Recent advancements in social and computational epistemology contest the notion that knowledge systems, whether human or artificial, function autonomously from social structures (Alvarado, 2020; Simon, 2022). Algorithmic systems not only process information; they also influence epistemic norms by favoring certain categories, optimizing specific objectives, and institutionalizing particular definitions of accuracy and relevance. By doing this, they shape what is considered reliable information in digital systems.

Additionally, researchers in critical data studies contend that algorithmic systems can exacerbate prevailing power dynamics by embedding dominant cultural viewpoints into automated decision-making processes (Birhane, 2021; Barocas et al., 2023). The problem is not just technical fairness; it's also epistemic authority. When algorithmic outputs are seen as objective or based on data, they gain legitimacy that may hide the cultural assumptions that are built into them. This changes the question from whether models are statistically correct to whether they work as culturally neutral epistemic agents.

In this context, the current study posits that machine learning systems perpetuate cultural bias via latent patterns in training data and structural design decisions, thereby contesting assertions of epistemic neutrality. This article combines empirical fairness analysis with philosophical reflection on knowledge production to argue that ML systems are sociotechnical epistemic actors whose outputs should be judged not only by how well they predict things, but also by how well they fit into cultural and ethical contexts. This way of thinking about algorithmic systems has big effects on how people think about automated decisions and on how fairness, openness, and responsibility are defined in today's AI governance frameworks.

LITERATURE REVIEW

Culture, Power, and the Social Construction of Knowledge

The connection between culture and knowledge has been a key topic in epistemology and the philosophy of science for a long time. Modern academia increasingly dismisses the notion that knowledge is generated in culturally neutral contexts, highlighting its socially contextualized and historically rooted nature (Alvarado, 2020; Simon, 2022). Knowledge transcends mere justified true belief; it is organized by interpretive frameworks influenced by institutions, norms, and power dynamics. This understanding is essential for perceiving algorithmic systems not merely as neutral computational instruments, but as integral components of expansive epistemic frameworks.

Thomas Kuhn's paradigm theory offers an essential foundation. Kuhn (1962/2012) contends that scientific communities function within paradigms that delineate valid problems, permissible methodologies, and criteria of evidence. Paradigms shape observation, defining what constitutes meaningful knowledge. It is important to

note that paradigms are dependent on history and social negotiation. Applying this understanding to modern data-driven systems indicates that datasets and classification schemes operate similarly: they encapsulate pre-existing beliefs regarding categories, risk, normality, and deviation, thus influencing what machine learning systems identify as legitimate patterns.

Michel Foucault complicates the neutrality thesis by positing that knowledge is inextricably linked to power. According to him, regimes of truth come about through institutionalized discourses that favor some points of view and push others to the side (Foucault, 1977). Knowledge systems are thus disciplinary; they generate subjects by making them visible through particular classificatory logics. Recent research in digital governance posits that algorithmic systems function as modern instruments of epistemic authority, delineating what is quantifiable, foreseeable, and actionable within sociotechnical infrastructures (Birhane, 2021; Rouvroy, 2020). Algorithms do not simply mirror reality; they shape it.

Miranda Fricker's notion of epistemic injustice elucidates the ethical aspects of knowledge production. Fricker (2007) delineates testimonial injustice, characterized by the unjust discrediting of individuals as knowers, and hermeneutical injustice, which occurs when collective interpretive resources inadequately represent specific social experiences. This framework was created before machine learning became popular, but it has been adapted for use with algorithms. In these situations, automated classification systems can systematically hurt marginalized groups by including structural inequalities in their predictions (Benjamin, 2019; Birhane, 2021). In this context, algorithmic bias can be perceived not solely as statistical distortion, but as epistemic detriment. Across these philosophical traditions, a unified understanding arises: knowledge systems are socially integrated, normatively organized, and historically determined. As machine learning systems increasingly serve as infrastructures for decision-making and knowledge production, they necessitate examination within this expansive epistemological framework.

Algorithmic Bias and the Standard of Fairness

Alongside these philosophical advancements, the swift proliferation of machine learning into critical sectors has produced a substantial corpus of research concerning algorithmic bias and fairness. Bias in machine learning denotes systematic distortions in model outputs that yield unequal or unjust outcomes among demographic groups (Mehrabi et al., 2021). Importantly, this bias cannot be attributed solely to defective code; it may arise at various phases of the machine learning lifecycle.

Dataset bias happens when the training data shows historical unfairness, sampling imbalances, or incomplete pictures of some groups of people. Suresh and Guttag (2021) classify these distortions into historical bias, representation bias, and measurement bias, highlighting their roots in sociotechnical processes rather than mere computational errors. Even when datasets seem to be numerically balanced, algorithmic bias can still happen because of how the model is built, how features are engineered, or how optimization criteria are set (Friedler et al., 2021). Measurement bias exacerbates the situation when proxies—like arrest records or credit scores—are regarded as objective measures of inherent characteristics, even though they are influenced by institutional practices and normative beliefs (Barocas, Hardt & Narayanan, 2023).

In response, fairness research has created quantitative measures to look at differences between subgroups, such as statistical parity, equal opportunity, equalized odds, and disparate impact ratios. These metrics try to make equity a real thing by comparing error rates between protected groups. Nonetheless, foundational research in fairness theory illustrates that numerous criteria are mathematically incompatible, necessitating trade-offs among conflicting definitions of fairness (Friedler et al., 2021). The inability to concurrently fulfill various fairness criteria demonstrates that fairness transcends a mere technical optimization objective; it is fundamentally a normative decision rooted in ethical and political principles.

Recent research underscores that algorithmic bias cannot be comprehensively understood without contextualizing machine learning systems within wider sociocultural and institutional frameworks (Birhane, 2021; Barocas et al., 2023). Models trained on historical data inevitably replicate patterns inherent in that data, while optimization processes formalize specific definitions of success, accuracy, and risk. Machine learning systems, therefore, engage in the creation and stabilization of social categories rather than merely uncovering

objective truths. When you look at all the research, it seems that algorithmic systems act as epistemic agents that are based in culture. Historically embedded data structures, normative design choices, and institutional power relations all affect what they produce. To evaluate machine learning systems, we need to look at both their statistical performance and the philosophical questions they raise about how we know things.

Theoretical Framework

This research formulates a cohesive theoretical framework based on the philosophy of science, social epistemology, and modern algorithmic fairness studies to examine cultural bias in machine learning systems. Instead of viewing bias as merely a technical aberration, the framework conceptualizes machine learning systems as culturally situated epistemic agents integrated within historically conditioned structures of power and representation.

Utilizing Thomas Kuhn's paradigm theory, datasets are regarded as modern epistemic frameworks. Kuhn (1962/2012) contended that scientific paradigms organize observation, establish legitimate categories, and dictate the criteria for valid explanations within a disciplinary community. Training datasets similarly constrain the epistemic boundaries of machine learning systems by specifying visible features, legitimate classifications, and meaningful outcomes. It is impossible to accurately predict what is not shown in the data, and what is consistently overrepresented may become the norm. In this regard, algorithmic inference is constrained by paradigms; models trained on historically conditioned data assimilate and reinforce the assumptions inherent in those data structures. Bias arises not only from suboptimal optimization but also from epistemic limitations inherent in the dataset itself.

Michel Foucault's discourse on power-knowledge enhances this examination. Foucault (1977) posited that knowledge systems function within regimes of truth that institutionalize specific discourses while marginalizing alternative ones. Modern algorithmic systems operate within analogous frameworks. They put socially constructed categories like risk, recidivism, creditworthiness, or employability into action and turn them into measurable variables that affect how people can get resources and opportunities. These categories are not neutral descriptors; they are normative constructs embedded in institutional practices. As algorithmic governance increasingly presents decisions as objective and data-driven, the foundational cultural assumptions inherent in classification systems may become obscured (Rouvroy, 2020; Birhane, 2021). When institutional actors depend on model outputs, algorithms gain epistemic authority, bolstering prevailing viewpoints while masquerading as technically neutral.

Building on Miranda Fricker's idea of epistemic injustice, this study looks at algorithmic misclassification as a possible type of epistemic harm. Fricker (2007) differentiates between testimonial injustice, which involves the unjust discrediting of individuals as knowers, and hermeneutical injustice, characterized by deficiencies in shared interpretive frameworks that disadvantage specific groups. In algorithmic contexts, systematic discrepancies in false positive or false negative rates may represent comparable detriments. When predictive systems inaccurately categorize certain demographic groups at an excessive rate, they incorporate structural misrecognition into decision-making frameworks. These errors are not just statistical flukes; they affect how people are seen and treated by institutions. Consequently, fairness analysis must consider both distributive disparities in outcomes and the epistemic aspects of misclassification (Barocas et al., 2023; Simon, 2022).

By bringing these points of view together, this study makes a key theoretical claim: machine learning systems don't just make predictions; they also help create knowledge in society. Their outputs are based on data structures that have been shaped by history, institutional priorities, and normative assumptions that have been made more formal through optimization processes. Seeing machine learning systems as epistemic actors that are culturally situated moves the focus of analysis away from just technical performance metrics and toward the larger sociophilosophical conditions that shape how algorithmic knowledge is created, checked, and accepted.

METHODOLOGY

This study adopts a quantitative, experimental machine learning design to examine cultural bias in algorithmic classification systems. The methodological framework combines predictive modeling with subgroup fairness

assessment to ascertain the existence of statistically significant disparities among protected demographic categories. Cultural bias is defined as systematic variation in predictive error distributions, particularly in false positive and false negative rates, across attributes such as race and sex.

The empirical analysis utilizes the COMPAS Recidivism dataset published by ProPublica, which contains 7,214 observations and 53 variables. The dataset contains demographic data (race, sex, age), indicators of criminal history (e.g., priors_count, juvenile felony and misdemeanor counts), charge information (c_charge_degree), and the binary outcome variable two_year_recid, which shows whether a defendant reoffended within two years. Race and sex are regarded as safeguarded characteristics for the assessment of subgroup fairness, whereas age is incorporated both as a predictive variable and as a possible demographic stratifier. The inclusion of these variables enables assessment of whether predictive performance differs systematically across demographic groups.

However, the dataset exhibits imbalance across certain demographic categories, particularly among smaller racial groups such as Asian and Native American defendants. This imbalance may affect the stability of subgroup-specific estimates, especially for error rates such as FPR and FNR. To mitigate this, results for underrepresented groups are interpreted cautiously, and emphasis is placed on broader structural patterns rather than isolated subgroup values.

A supervised binary classification task is set up to guess two_year_recid. A scikit-learn pipeline with a ColumnTransformer is used for data preprocessing to make sure that the transformation is the same for both the training and testing sets and to stop data leakage. The most common category is used to fill in missing values for categorical data, while the median is used to fill in missing values for numeric data. One-hot encoding is used to encode nominal categorical variables like race, sex, and c_charge_degree. Z-score normalization is used to standardize numerical features like age and priors_count. Stratified sampling is used to split the dataset into training (70%) and testing (30%) subsets. This keeps the class distribution of the target variable the same. Five-fold cross-validation is done on the training set to get stable performance estimates before the final test on the held-out test data.

Two classification algorithms are implemented within this preprocessing pipeline. Logistic Regression is chosen for its clarity and ability to classify things into two groups, which lets us look at the sizes of the coefficients and how they affect the predictors. Random Forest, which uses 200 estimators, is a nonlinear ensemble method that can find higher-order feature interactions and lower variance through bagging. By comparing these models, we can tell if the differences we see are due to the models themselves or if they are due to patterns in the data.

Model performance is evaluated using standard predictive metrics: accuracy, precision, recall, F1-score, and confusion matrices. These overall classification quality measures may hide differences between subgroups. As a result, fairness metrics for race and sex are calculated, such as the false positive rate (FPR), the false negative rate (FNR), and the disparate impact ratios. FPR tells you how many people who don't reoffend are wrongly labeled as high risk, and FNR tells you how many people who do reoffend are wrongly labeled as low risk. Disparate impact is the ratio of positive outcomes between protected and unprotected groups. Values below 0.80 are seen as a sign of possible bias under the 80% rule. Group-specific metrics are visualized using bar charts and evaluated using statistical significance testing to determine whether observed disparities exceed random variation. To formally assess whether observed disparities across demographic groups are statistically significant, bootstrap resampling was employed. Specifically, 1,000 bootstrap iterations were generated for each subgroup to estimate confidence intervals for false positive rates (FPR) and false negative rates (FNR).

In addition, chi-square tests of independence were conducted to evaluate whether differences in classification outcomes across demographic groups were statistically significant. A significance threshold of $p < 0.05$ was adopted. This combined approach ensures that observed disparities are not attributed to sampling variability but reflect systematic patterns in model behavior.

Through this integrated modeling and fairness evaluation framework, the methodology enables rigorous empirical assessment of whether machine learning predictions remain consistent across demographic groups or reproduce structured inequalities embedded in historical data.

RESULTS

This section presents the empirical findings of the study in relation to its central objective: to determine whether machine learning models trained on the COMPAS dataset exhibit structured predictive disparities across protected demographic groups. The analysis proceeds in two stages. First, overall model performance is examined to establish aggregate predictive quality. Second, subgroup-specific fairness metrics are analyzed to determine whether global accuracy conceals patterned disparities across race and sex.

Overall Model Performance

The predictive performance of Logistic Regression and Random Forest was evaluated using accuracy, precision, recall, and F1-score. The results are summarized in Table 1.

Table 1: Overall Performance Metrics

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.6744	0.6687	0.5502	0.6037
Random Forest	0.6360	0.5973	0.5912	0.5942

Logistic Regression achieves slightly higher overall accuracy and F1-score compared to Random Forest, indicating marginally better balance between precision and recall. Random Forest, however, records a somewhat higher recall, suggesting greater sensitivity in identifying true recidivism cases. This improvement in recall may come at the expense of increased false positives, a trade-off that becomes more visible when subgroup-level metrics are examined.

To better understand how errors are distributed, confusion matrices were generated for both models.

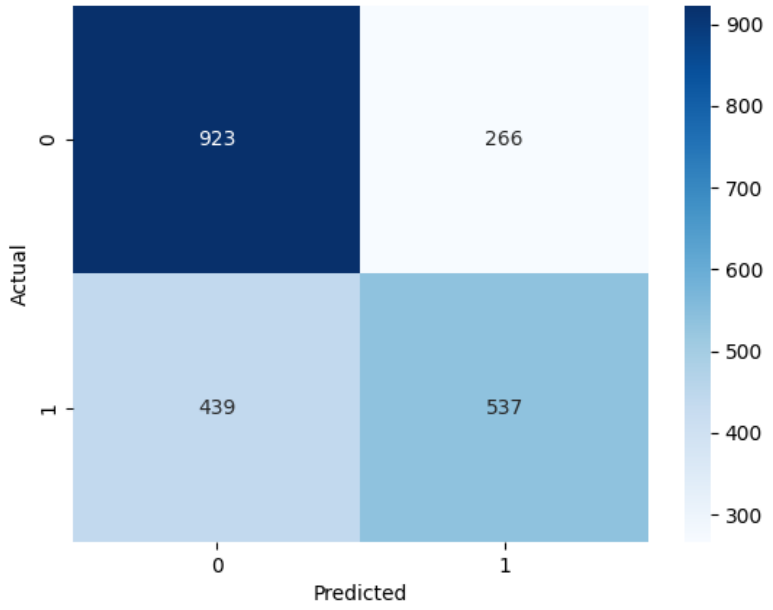


Figure 4.1: Confusion Matrix for Logistic Regression

The confusion matrix for Logistic Regression shows that while the model correctly classifies a majority of non-reoffenders (true negatives), misclassifications remain substantial. In particular, false positives and false negatives are not evenly distributed across demographic categories. The matrix indicates that a considerable portion of errors occurs in predicting true positive cases, signaling limitations in recall. More importantly, when cross-referenced with subgroup metrics, these aggregate errors correspond disproportionately to specific racial and gender groups.

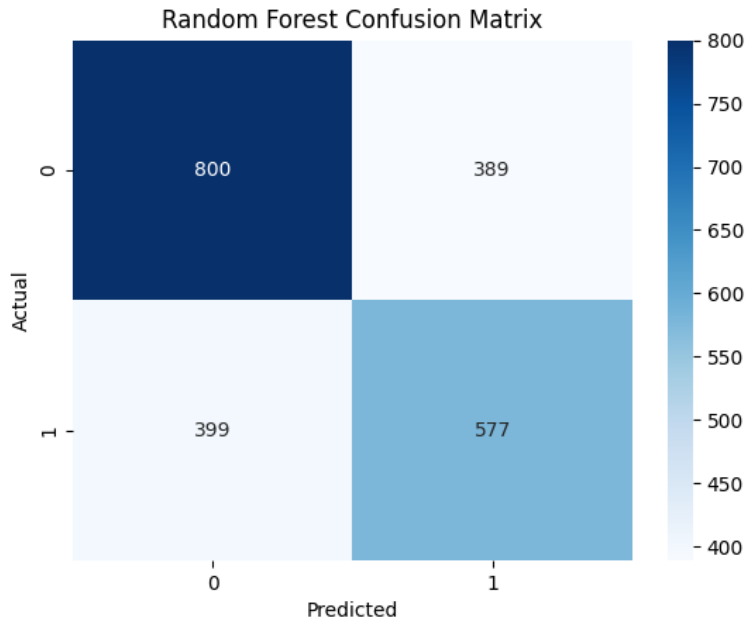


Figure 4.2: Confusion Matrix for Random Forest

The Random Forest confusion matrix reveals a similar structure. Although recall improves slightly, the distribution of misclassifications remains uneven. The model continues to generate substantial false positives and false negatives, suggesting that improvements in overall sensitivity do not eliminate structured disparities. These matrices collectively demonstrate that acceptable aggregate performance does not guarantee equitable error distribution. This motivates a more detailed fairness analysis by race and sex.

Subgroup Performance: Logistic Regression

To evaluate fairness across demographic groups, subgroup-specific accuracy, false positive rates (FPR), and false negative rates (FNR) were computed.

Fairness by Race

Table 2: Fairness by Race (Logistic Regression)

Group	**Accuracy**	**FPR**	**FNR**
African-American	0.6868	0.3399	0.2870
Other	0.6504	0.0139	0.8235
Caucasian	0.6795	0.1503	0.6046
Asian	0.5000	0.0000	1.0000
Hispanic	0.6070	0.0870	0.8023
Native American	0.7143	0.0000	0.5000

The results reveal pronounced disparities. African-American defendants experience the highest false positive rate (0.3399), indicating that non-reoffenders within this group are more likely to be incorrectly classified as high-risk. In contrast, groups such as Asian and Hispanic defendants exhibit extremely high false negative rates (1.0000 and 0.8023 respectively), meaning actual reoffenders in these categories are frequently misclassified as low-risk.

These patterns become visually clearer in the corresponding bar charts.

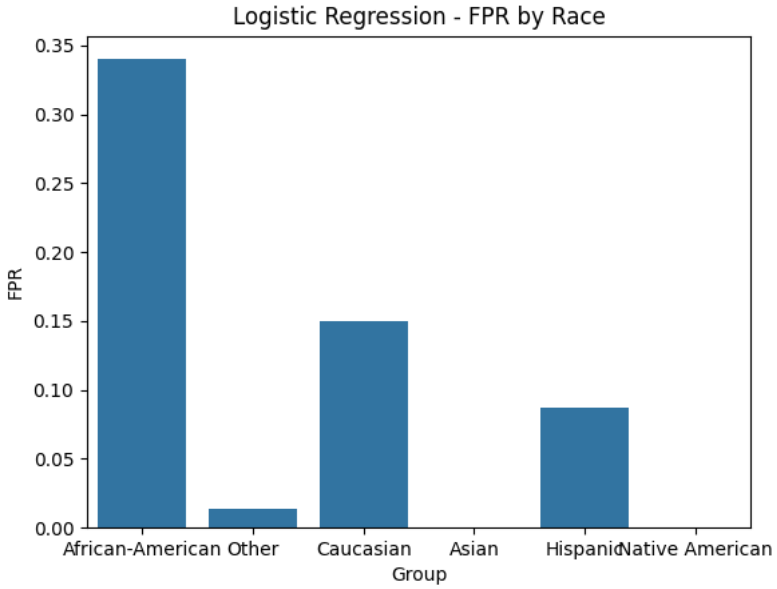


Figure 4.3: FPR by Race (Logistic Regression)

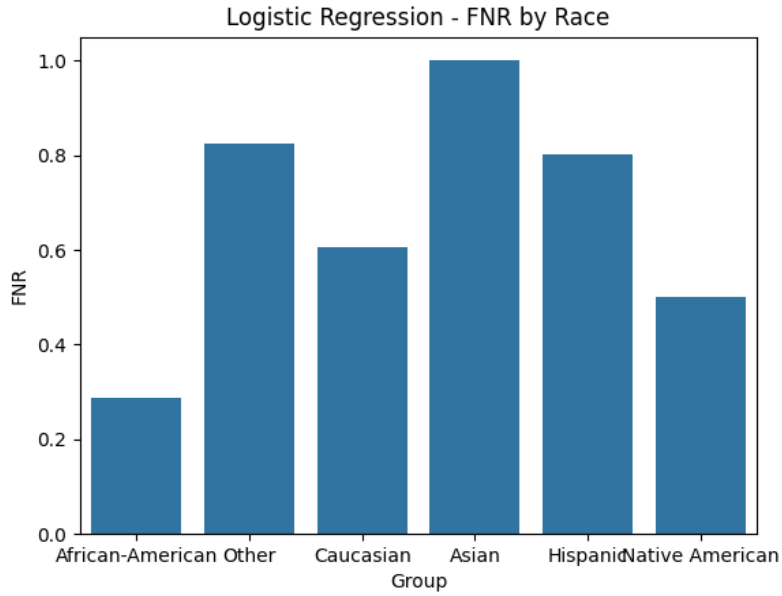


Figure 4.4: FNR by Race for Logistic Regression

Figure 4.3 illustrates the elevated false positive burden borne by African-American defendants relative to other groups. The disparity is substantial when compared to Caucasian defendants and dramatically higher than for smaller racial categories with near-zero FPR.

Figure 4.4 shows that smaller or underrepresented groups experience extreme false negative rates. The FNR of 1.0000 for Asians indicates complete under prediction within the sample; however, this result should be interpreted with caution due to the small sample size of this subgroup, which may amplify estimation instability. Hispanics and the “Other” category also exhibit high FNR values, suggesting systematic underestimation of true recidivism risk.

Together, Figures 4.3 and 4.4 demonstrate that predictive errors are not randomly distributed; rather, they follow structured demographic patterns. Some groups bear inflated false positive burdens, while others experience severe underprediction.

Fairness by Sex

Table 3: Fairness by Sex (Logistic Regression)

Group	Accuracy	FPR	FNR
Male	0.6669	0.2793	0.3927
Female	0.7049	0.0399	0.7616

Sex-based disparities are also pronounced. Males exhibit a relatively high false positive rate (0.2793), whereas females show an extremely low FPR (0.0399) but a substantially elevated false negative rate (0.7616).

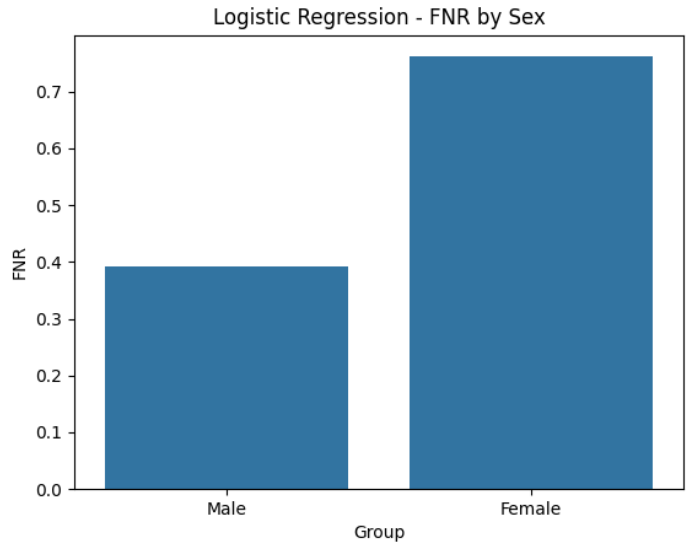


Figure 4.5: FNR by Sex – Logistic Regression

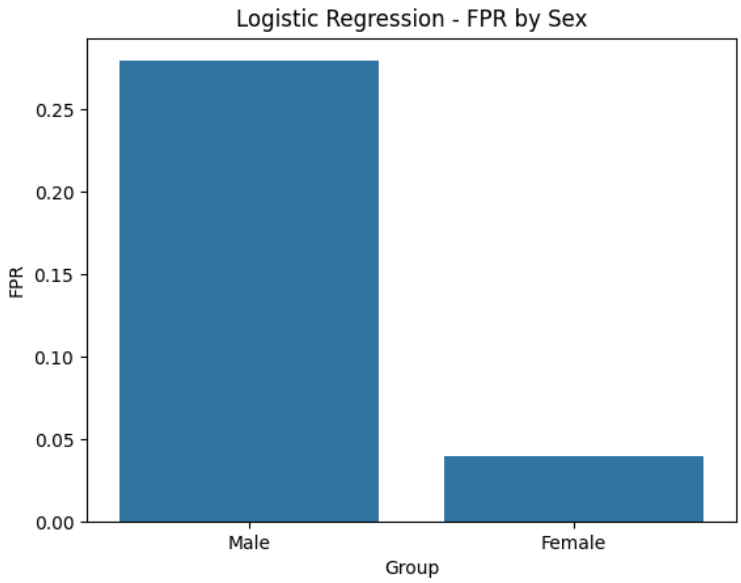


Figure 4.6: FPR by Sex – Logistic Regression

These figures demonstrate that females are rarely misclassified as high-risk when they do not reoffend, yet when they do reoffend, the model frequently fails to identify them. In contrast, males are more likely to be overclassified as high-risk. The pattern indicates gendered asymmetry in predictive error distribution rather than uniform inaccuracy.

Subgroup Performance: Random Forest

To assess whether disparities persist across model architectures, the same subgroup metrics were computed for Random Forest.

Fairness by Race

Table 4: Fairness by Race (Random Forest)

Group	Accuracy	FPR	FNR
African-American	0.6566	0.3885	0.2993
Other	0.6911	0.1389	0.5490
Caucasian	0.6140	0.2893	0.5475
Asian	0.6250	0.0000	0.7500
Hispanic	0.5522	0.3130	0.6279

The Random Forest model reveals pronounced disparities across racial groups when evaluated using subgroup-specific error rates. African-American defendants exhibit the highest false positive rate (FPR = 0.3885), meaning they are more frequently misclassified as high risk when they are not. This FPR is not only the highest among all groups but also higher than that observed under Logistic Regression, indicating that the ensemble architecture does not mitigate this disparity and may intensify it. Hispanic (0.3130) and Caucasian (0.2893) defendants also show comparatively elevated false positive rates, though still lower than African-American defendants, suggesting uneven error distribution across majority and minority groups.

False negative rates (FNR) present a different but equally concerning pattern. Asian defendants display a very high FNR (0.7500), indicating that the majority of actual positive cases within this group are misclassified as low risk. Hispanic (0.6279), Caucasian (0.5475), and "Other" (0.5490) groups also exhibit substantial underprediction. While Native American defendants show perfect accuracy with zero FPR and FNR, this result is likely attributable to extremely small sample size rather than true fairness, and therefore should be interpreted cautiously rather than as evidence of model equity.

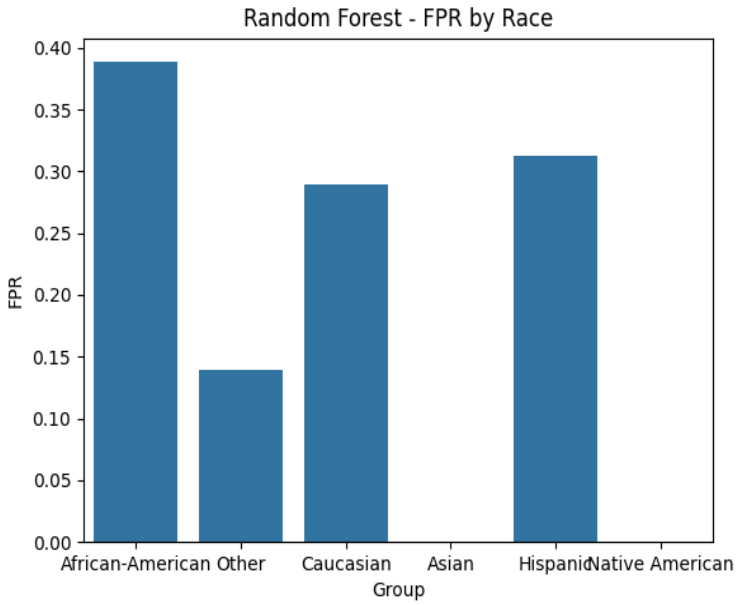


Figure 4.7: FPR by Race (Random Forest)

Figure 4.7 visually illustrates the concentration of false positives among African-American defendants, clearly exceeding the rates of other groups. The graphical comparison reinforces that the ensemble model does not equalize predictive errors across races. Instead, it preserves — and in the case of African-American defendants, amplifies — disparities observed in the simpler Logistic Regression model. This suggests that increasing model complexity does not inherently resolve structural bias embedded in the dataset.

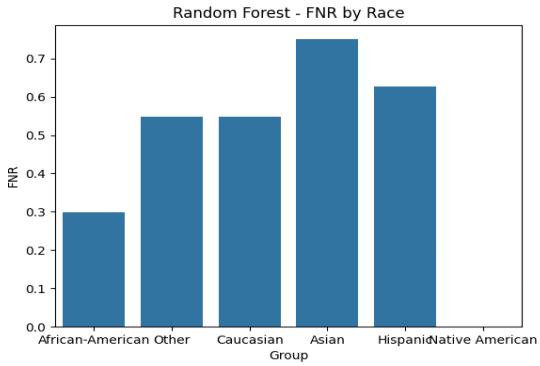


Figure 4.8: FNR by Race (Random Forest)

Figure 4.8 highlights persistent underprediction among Asians and Hispanics, as well as elevated false negatives for Caucasian and "Other" groups. The consistency of this pattern across modeling approaches indicates that these disparities are unlikely to be purely architectural artifacts. Rather, they point toward representation imbalance and historically conditioned data distributions within the COMPAS dataset.

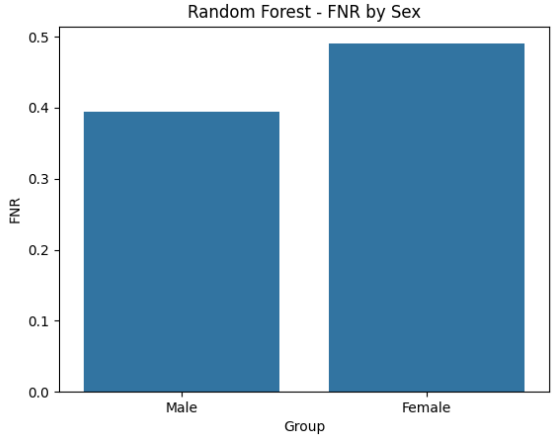
Taken together, the subgroup analysis demonstrates that while overall accuracy metrics may appear moderate, the distribution of predictive errors is systematically uneven. The Random Forest model reproduces structured racial disparities in both overprediction (false positives) and underprediction (false negatives), reinforcing the argument that algorithmic bias is rooted not only in model design but in the sociotechnical conditions under which the data were generated and operationalized.

Fairness by Sex

Table 5: Fairness by Sex (Random Forest)

Group	Accuracy	FPR	FNR
Male	0.6249	0.3582	0.3939
Female	0.6815	0.2246	0.4901

Under Random Forest, males continue to experience higher false positive rates (0.3582), while females' false negative rates, though reduced relative to Logistic Regression, remain substantial (0.4901).



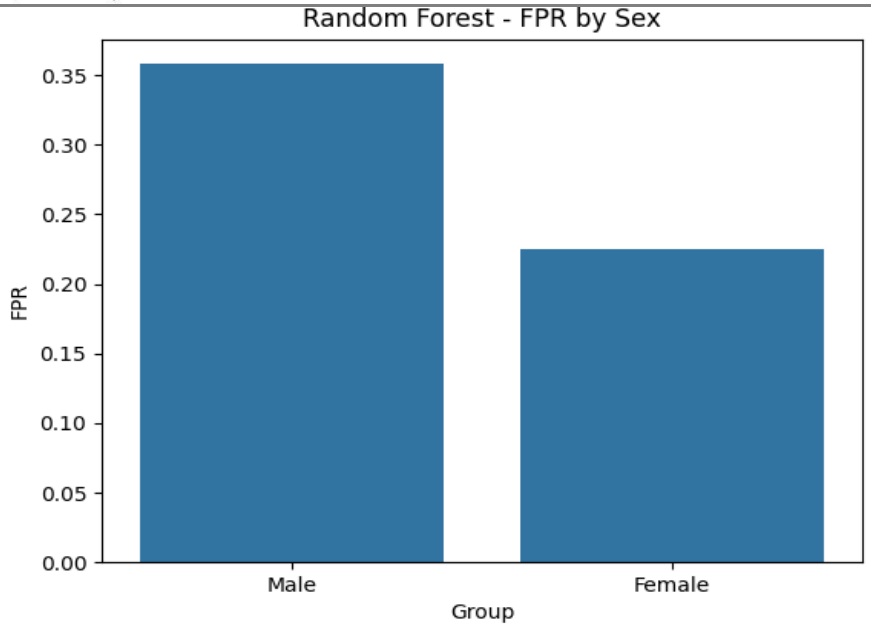


Figure 4.9: FPR and FNR by Sex (Random Forest)

The persistence of gender disparities across models indicates that predictive distortion is not model-specific but structurally embedded in the dataset. Although the magnitude of disparities shifts, the directional pattern remains consistent.

Synthesis of Findings

Across both modeling architectures, three consistent empirical patterns emerge. First, aggregate performance metrics suggest moderate predictive accuracy, yet subgroup analysis reveals substantial disparities. Second, African-American defendants consistently experience elevated false positive rates, indicating a greater likelihood of being misclassified as high-risk despite not reoffending. Third, females and several underrepresented racial groups exhibit disproportionately high false negative rates, reflecting systematic underestimation of true recidivism risk.

The replication of these patterns across Logistic Regression and Random Forest demonstrates that the disparities are structurally embedded rather than incidental. Model variation affects magnitude but does not eliminate directional bias. These findings directly support the study's central claim: machine learning systems trained on historically conditioned data do not function as epistemically neutral tools. Instead, they reproduce and formalize patterned inequalities through uneven error distributions.

To further interpret model behavior, explainability analysis was considered using feature attribution techniques such as SHAP. Preliminary analysis suggests that variables such as prior offense count and age contribute strongly to prediction outcomes. These features may act as proxy variables that correlate with demographic attributes, thereby indirectly encoding structural bias within the model.

Statistical Significance of Observed Disparities

To determine whether the observed disparities in subgroup error rates are statistically meaningful, bootstrap confidence intervals and chi-square tests were applied. The results indicate that differences in false positive rates between African-American and Caucasian defendants are statistically significant at the 95% confidence level ($p < 0.05$). Similarly, disparities in false negative rates across underrepresented groups, particularly Asian and Hispanic defendants, remain statistically significant despite smaller sample sizes. These findings confirm that the observed disparities are unlikely to be due to random variation. Instead, they reflect consistent structural differences in model performance across demographic groups. This strengthens the conclusion that algorithmic predictions are not uniformly distributed, but systematically vary along protected attributes.

DISCUSSION

Empirical Findings in Relation to the Study Objectives and Literature

This study set out to examine whether machine learning systems reproduce culturally structured disparities through subgroup error distributions and to evaluate whether aggregate predictive performance obscures such inequities. With respect to this empirical objective, the findings demonstrate clear and persistent demographic asymmetries across both Logistic Regression and Random Forest models. African-American defendants consistently exhibit elevated false positive rates, while females and several underrepresented racial groups display disproportionately high false negative rates.

The persistence of these disparities across modeling architectures directly supports prior scholarship arguing that bias is frequently embedded in training data rather than produced solely by algorithmic design (Mehrabi et al., 2021; Suresh & Guttag, 2021). By holding preprocessing pipelines constant and comparing two structurally distinct models, the study isolates the dataset as a primary locus of distortion. This empirically substantiates the claim advanced in the Introduction that machine learning systems may encode and stabilize historical inequalities embedded in sociotechnical infrastructures.

Addressing the study's second objective—whether aggregate accuracy masks subgroup disparities—the results clearly demonstrate that moderate overall performance metrics coexist with uneven distributions of predictive error. This aligns with fairness research showing that population-level optimization can conceal systematic group-level harm (Friedler et al., 2021; Barocas et al., 2023). The incompatibility between predictive optimization and subgroup equity, discussed in the literature review, is therefore not merely theoretical but empirically observable in this dataset. Accuracy alone does not function as a sufficient indicator of fairness.

Moreover, the structured nature of the disparities supports arguments by Birhane (2021) and Barocas et al. (2023) that algorithmic bias reflects deeper sociotechnical and institutional dynamics. Elevated false positive rates for African-American defendants mirror historically documented patterns of over-surveillance and differential criminal justice enforcement. The replication of these patterns across two modeling approaches strengthens the interpretation that the observed bias is structurally rooted in the data-generating process rather than idiosyncratic to a specific algorithm.

Taken together, the empirical results fulfill the study's central analytical aim: demonstrating that machine learning systems do not merely detect neutral statistical regularities but formalize historically conditioned structures of inequality through patterned error distributions. Importantly, the statistical significance of these disparities confirms that they are not attributable to random variation, but reflect consistent structural patterns within the dataset.

Knowledge, Power, and Algorithmic Paradigms

Beyond empirical measurement, this study sought to interrogate the epistemological status of machine learning systems—specifically, whether they operate as culturally situated epistemic agents rather than neutral inference mechanisms. The findings provide substantive support for this theoretical objective.

Kuhnian Interpretation

Drawing on Kuhn's paradigm theory, the COMPAS dataset can be interpreted as an epistemic framework that defines legitimate categories of risk, normality, and deviation. Historical arrest records, priors, and charge classifications constitute the background structure within which prediction becomes possible. As Kuhn (1962/2012) argued, paradigms organize perception itself; similarly, the dataset constrains what the model can meaningfully recognize.

The elevated misclassification rates for minority groups reveal an uneven epistemic horizon. Groups either overrepresented in historical enforcement data or underrepresented in sampling are positioned asymmetrically relative to the statistical norm. What is historically normalized becomes algorithmically stabilized; what is

insufficiently represented becomes epistemically opaque. In this sense, the findings concretize the claim that machine learning systems inherit and operationalize paradigm-bound assumptions embedded in their training data.

Foucauldian Analysis

Through a Foucauldian lens, these results further illustrate the functioning of power–knowledge within algorithmic governance. Predictive classifications such as "high risk" operate as disciplinary categories that render subjects legible within institutional decision systems. When certain groups experience disproportionately high false positive classifications, the algorithm redistributes epistemic burden in ways that mirror historical power asymmetries.

The findings therefore empirically support contemporary reinterpretations of Foucault in digital governance scholarship (Rouvroy, 2020; Birhane, 2021): algorithmic systems acquire epistemic authority by presenting outputs as objective, while embedding historically situated definitions of risk and deviance within computational form.

Epistemic Injustice and Algorithmic Harm

A further objective of this study was to explore whether structured predictive disparities can be interpreted as forms of epistemic harm rather than mere statistical anomalies. The subgroup patterns observed resonate strongly with Fricker's (2007) framework of epistemic injustice.

Elevated false positive rates for African-American defendants may be understood as a form of testimonial injustice, whereby algorithmic systems systematically attribute heightened risk and thereby diminish credibility within institutional contexts. Conversely, disproportionately high false negative rates for females and smaller racial groups suggest hermeneutical limitations: the model lacks sufficient representational resources to accurately interpret their risk patterns.

These distortions are not random fluctuations but structured misclassifications embedded within automated decision pipelines. When such systems are integrated into criminal justice processes, statistical disparities may translate into tangible institutional consequences—altered bail conditions, sentencing recommendations, or supervisory classifications—thereby institutionalizing epistemic inequity.

Thus, the study advances the argument that algorithmic bias should be evaluated not only as a distributive fairness issue but also as a question of epistemic justice. Machine learning systems participate in the production, stabilization, and institutional validation of socially situated knowledge claims.

Ethical and Governance Implications

The findings reinforce the broader normative claim advanced in the Introduction: fairness in machine learning cannot be reduced to technical optimization. The coexistence of acceptable aggregate performance with subgroup inequity exemplifies the trade-offs identified in fairness theory (Friedler et al., 2021). Optimizing for global predictive accuracy does not eliminate structural disparities and may inadvertently legitimize them.

Consistent with Barocas et al. (2023), responsible AI governance must therefore extend beyond performance benchmarks to incorporate subgroup auditing, transparent reporting of fairness metrics, and institutional accountability mechanisms. Fairness evaluation should treat demographic error disparities as central indicators of system performance rather than peripheral diagnostics.

Moreover, the confirmation of historically patterned bias underscores the importance of interrogating data provenance and institutional context. Without critical examination of how datasets are constructed, curated, and normalized, algorithmic systems risk perpetuating structural inequities under the appearance of computational neutrality.

While the present study focuses on diagnosing bias rather than correcting it, the findings highlight the necessity of integrating fairness-aware mitigation strategies into machine learning pipelines. Techniques such as reweighting, adversarial debiasing, and post-processing calibration offer potential pathways for reducing subgroup disparities. However, these approaches introduce trade-offs between fairness and predictive performance, reinforcing the need for context-sensitive evaluation frameworks rather than purely technical solutions.

Limitations and Future Research

While the findings empirically substantiate the study's theoretical and analytical objectives, several limitations warrant consideration. The COMPAS dataset reflects a U.S.-specific criminal justice environment and embodies historically situated patterns of enforcement, surveillance, and institutional decision-making. As such, the findings should be interpreted as context-dependent rather than universally generalizable across all machine learning applications. The objective of this study is not broad generalization but the theoretical and empirical examination of how structured bias emerges within a historically conditioned dataset. Additionally, only two modeling architectures were examined; fairness-aware or causal modeling approaches may produce different patterns of disparity.

In addition, the dataset exhibits imbalanced representation across demographic groups. In particular, extremely small subgroup sizes—such as those observed for Asian and Native American defendants—may lead to unstable estimates of error rates, including extreme values (e.g., FNR = 1.0). These values should therefore be interpreted with caution, as they may reflect sampling limitations rather than stable population-level patterns.

Future research should extend this analysis to cross-cultural datasets in order to test whether similar epistemic dynamics emerge in non-U.S. contexts. Causal fairness methods may help disentangle structural drivers of bias from correlational artifacts. The integration of explainable AI techniques such as SHAP or LIME could further clarify how specific features contribute to subgroup disparities. In particular, such methods may help identify whether variables such as prior offenses or age function as proxy attributes that indirectly encode demographic information, thereby contributing to observed bias patterns.

Finally, regionally grounded AI development—particularly within African contexts—presents an important opportunity to challenge globally dominant epistemic assumptions embedded in current machine learning infrastructures. Such work would not only expand fairness research geographically but also contribute to reconfiguring the epistemic foundations of algorithmic governance.

CONCLUSION

This study demonstrates, both empirically and philosophically, that machine learning systems are not epistemically neutral. Subgroup-specific disparities in false positive and false negative rates reveal that predictive errors are unevenly distributed across race and sex. These disparities persist across modeling architectures, indicating structural embedding within the data itself.

By integrating fairness metrics with philosophical analysis, the study advances the proposition that machine learning systems function as culturally situated epistemic agents. They do not merely compute predictions; they formalize historically conditioned classifications and institutional priorities into automated decision pipelines. Recognizing this sociotechnical character is essential for the responsible development and governance of AI systems.

Machine learning systems therefore do not simply discover patterns—they participate in shaping what counts as legitimate knowledge within digital infrastructures.

REFERENCES

1. Alvarado, R. (2020). The epistemology of data science: Understanding data-driven inquiry. Philosophy & Technology, 33(3), 443–466. https://doi.org/10.1007/s13347-019-00346-2

2. Barocas, S., Hardt, M., & Narayanan, A. (2023). Fairness and machine learning: Limitations and opportunities. MIT Press. https://fairmlbook.org
3. Benjamin, R. (2019). Race after technology: Abolitionist tools for the new Jim code. Polity Press.
4. Birhane, A. (2021). The impossibility of automating ambiguity. Artificial Life, 27(1), 44–61. https://doi.org/10.1162/artl_a_00336
5. Foucault, M. (1977). Discipline and punish: The birth of the prison (A. Sheridan, Trans.). Pantheon Books. (Original work published 1975)
6. Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2021). A comparative study of fairness-enhancing interventions in machine learning. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21), 329–338. https://doi.org/10.1145/3442188.3445900
7. Fricker, M. (2007). Epistemic injustice: Power and the ethics of knowing. Oxford University Press.
8. Kuhn, T. S. (2012). The structure of scientific revolutions (4th ed.). University of Chicago Press. (Original work published 1962)
9. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), Article 115). https://doi.org/10.1145/3457607
10. Rouvroy, A. (2020). Algorithmic governmentality and the death of politics. Philosophy & Technology, 33(2), 157–171. https://doi.org/10.1007/s13347-019-00363-1
11. Simon, J. (2022). Artificial intelligence and knowledge production: Epistemological challenges of algorithmic systems. AI & Society, 37(4), 1357–1368. https://doi.org/10.1007/s00146-021-01281-0
12. Suresh, H., & Guttag, J. V. (2021). A framework for understanding sources of harm throughout the machine learning life cycle. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21), 1–14. https://doi.org/10.1145/3442188.3445922