SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

Regulating Artificial Intelligence in the Age of Autonomous Decision-Making: A Legal and Ethical Framework

Dr. Alok Kumar Bhargava[1*], Dr. Sudarshan Nimma[2], Dr. Sourabh V. C. Ubale[3], Dr Jitendra Yadav[4], Dr Saleena Kuzhuppil Basheer[5], Dr Sidhartha Sekhar Dash[6]

[1]Founder and Principal Researcher, TrayiVani Foundation, Independent Research Organisation, TrayiVāṇī Eternal Verses on Peace, Silence & Discernment; Saviour Green Isle, Crossing Republik, Ghaziabad – 201016, Uttar Pradesh, India. E-mail: founder@trayivani.in, ORCID ID: 0009-0009-1805-3075
[2]Associate Professor, College of Law, Kakatiya University, Telangana. E-mail: snimma26@gmail.com
[3]Assistant Professor (PG), LL.M.-International Law, NET-JRF, PhD, D.Litt., Post-Doc, MM Shankarrao Chavan Law College, Pune (Permanently affiliated to Savitribai Phule Pune University), International Law, Technology Laws, Jurisprudence.
E-mail: ubale.sourabh@gmail.com, ORCID ID: 0000-0001-8371-4532
[4]Associate Professor, ILSR, Mangalayatan University, Beswan, Aligarh -202146. E-mail: jitendra.yadav@mangalayatan.edu.in
[5]PhD, Professor & Dean, Jamia Hamdard, Specialisation in law. E-mail: basheerk@jamiahamdard.ac.in, ORCID ID: 0009-0004-6277-691X
[6]Associate Professor, School of Law, KIIT Deemed to be University. E-mail: sidhartha@kls.ac.in

ABSTRACT
The growing use of autonomous systems that make decisions in high-stakes institutional contexts has raised growing concerns about predictive fairness, accountability and regulatory oversight. While it is often true that machine learning models improve classification accuracy, the distributive consequences of these models are controversial. In this study, the empirical performance and fairness implications of supervised learning models in the context of judicial recidivism risk assessment are empirically evaluated using a publicly available publicly available COMPAS recidivism dataset distributed via Kaggle. Linear and non-linear classifiers (Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine with RBF kernel) were benchmarked with respect to accuracy, ROC-AUC, Cross-Validation, Calibration Analysis and Subgroup Fairness Diagnostics. The Random Forest model showed a better discriminative power (AUC = 0.9065) and stable cross-validated generalization performance; however, the differences in the false positive rates of the different demographic groups were not negligible in spite of high performance in terms of accuracy aggregate. Statistical validation was done to ensure that difference in performance was systematic and not stochastic. These results show how there is a structural tension between predictive optimization and distributive equity, which point to the fact that technical improvements in performance cannot address questions of fairness. The study is consistent with risk-based approaches to regulation that integrate techniques of statistical validation, fairness audits and institutional accountability mechanisms. Effective management of autonomous AI systems therefore requires a balancing act of the predictiveness of the AI system and an equitable impact to foster legitimacy in high stakes decision environments.

Keywords: Autonomous decision-making; AI regulation; Algorithmic fairness; Recidivism prediction; Random Forest; Risk assessment; AI governance; Distributive equity

INTRODUCTION

The integration of artificial intelligence (AI) in institutional decision-making processes has been adopted at a rapid pace and this has brought about a change in the architecture of the governance process, be it that of a judicial, financial, healthcare or administrative. Contemporary machine learning systems are now playing a role in decisions previously made by the human aspect of discretion - bail decisions, sentencing recommendations, credit scoring, and predictive policing. These systems are based on statistical frameworks for learning that can describe complex non-linear relationship in large sets of data and quite often surpass the traditional linear ones in predictors [1]. While the technical sophistication of ensemble and supervised learning techniques have increased their predictive capacity, as they become more autonomous in a high-stake environment, there are foundational legal and ethical issues.

The development of systems that make decisions independently has heightened discussions about fairness, accountability and transparency. Early efforts in algorithmic fairness show that predictive accuracy maximization does not necessarily lead to fair treatment of demographic groups [2]. Moreover, theoretical results show that some fairness criteria are mutually incompatible when base rates are varying between subpopulations [3]. Empirical studies of recidivism risk assessment tools are just one more example of the

challenges that algorithmic systems, even when calibrated, can cause disparate error rates across racial groups [4]. The results are a challenge to the belief that the normative issues in automated governance can be solved by technical optimization.

Scholarship on law has also been drawing attention to the danger of data driven systems leading to disparate impact without the intent of discrimination [5]. High-profile analytical investigations have made the public sensitive to the problem of algorithmic discrimination in criminal justice settings, where it has been found that there are measurable racial disparities in predictive instruments [6]. This kind of evidence raises awareness of the conflict between predictive efficiency and distributive equity on the part of the autonomous systems. With the increased impact of algorithms on liberty restrictions decisions, the question of procedural fairness, due process and equal protection becomes pressing.

These developments have been met with the regulatory environment by coming up with new governance structures. Accountable algorithms are based on the notion that auditability, traceability, and institutional control are needed as key protective measures [7]. At the same time, criticisms of algorithmic opacity have described modern AI systems as "black boxes" with deficits of transparency in automated authority structures [8]. Ethical frameworks for AI governance emphasize the need for developing systems that are compatible with the principles of social good, such as fairness, non-maleficence, and explicability [9]. Nonetheless, the law on the so-called right to explanation shows that there is always ambiguity as to whether transparency requirements can be enforced and implemented in automated decision systems [10].

These principles have been attempted to be formalized by recent regulatory initiatives. The proposed Artificial intelligence act by the European Union proposes risk-based classification and compliance requirements of high-risk AI systems [11]. Complementarily, the United States National Institute of Standards and Technology (NIST) has created the Artificial Intelligence Risk Management Framework for the responsible development and deployment of AI [12]. These control tools indicate increasing awareness that predictive accuracy is not a sufficient validation measure of autonomous systems that act in socially consequential areas.

Although the regulatory discourse has been growing, one question that has not been answered is how can policymakers balance predictive efficiency and distributive justice in autonomous decision-making systems? While machine learning studies are still optimizing technical methodologies, in normative evaluation, empirical analysis is needed concerning the high-performance models distributed errors on the demographic groups. Unless there is a thorough analysis that combines performance measures with fairness diagnostic, regulatory frameworks are subject to either excessive estimation of the impartiality of automated systems or excessive underestimation of their predictive value.

This study addresses this lack of literature by empirically assessing supervised learning models in a criminal justice risk assessment context. Using ensemble and baseline classifiers the research compares, in a systematic way, predictive performance, calibration properties and subgroup error distributions. By combining statistical validation and fairness auditing, the research aims to take a closer look at the question of whether improved predictive discrimination necessarily alleviates or aggravates distributive disparities. The investigation puts the technical results in perspective of more general issues of legal and regulatory debate, and so plays the role of a bridge between computational modelling and regulatory theory.

The scope of the study is limited to structured recidivism data and the supervised classification methods. While this focus offers provision for the exact evaluation of the quantitative approach, it excludes the more general sociological determinants of the crime and the qualitative institutional factors that affect the discretion of the judiciary. Moreover, the concept of fairness is operationalized in terms of disparities in the error rates and calibration measures instead of counterfactual causal models. These limitations are associated with the fact that the study focuses on measurable predictive dynamics rather than thorough normative adjudication.

The importance of the research is that it combines empirical machine learning evaluation and regulation analysis. rather than addressing the concept of predictive accuracy and fairness in isolation, the research investigates the interplaying nature of the two in an autonomous decision-making paradigm. The results can be used to fuel the current discussions on AI governance, institutional responsibility, and algorithmic legitimacy by empirically proving the coexistence of high discriminative performance and quantifiable subgroup differences.

In light of the theoretical tensions between predictive optimization and distributive justice identified in prior scholarship and consistent with emerging regulatory requirements for high-risk AI systems also this study pursues the following research objectives:

- To empirically compare and contrast the predictive accuracy and cross-validation stability of the linear and nonlinear supervised learning models in a high-stakes autonomous decision-making scenario, the structural discriminative capacity is determined using the accuracy, ROC-AUC, precision-recall analysis, calibration performance, and cross-validation.
- To obtain a systematic evaluation of demographic fairness, based on analysis of the distributions of subgroup errors, such as false positive rates, true positive rates, and calibration differences, and thus to determine whether the improvements in predictive performance reduce or replicate the fairness Efficiency trades-offs in the computational fairness theory.
- To make the empirical results relevant to the modern AI governance and regulatory framework, assessing

the role of predictive superiority and perceived distributive inequalities on designing the accountability mechanisms, transparency demands, and risk-based regulatory supervision in autonomous decision systems.

2. Literature review

Self-regulation of autonomous decision-making systems has become the main issue of modern technology regulation. Since the use of artificial intelligence is becoming widespread in the context of high-risk institutional settings, regulatory frameworks have attempted to codify the concept of accountability, transparency, and risk reduction. The Artificial Intelligence Risk Management Framework of the National Institute of Standards and Technology focus on the systematic risk identification, measurement, and monitoring systems of AI systems used in consequential areas [13]. In its complement, interpretability scholarship holds that the high-stakes decision systems must have transparent and auditable logic in order to make the institution legit [14]. Nevertheless, the very idea of interpretability is debatable, and it has been argued that explainability is technically ambiguous and normatively weak when deinstitutionalized [15]. New attempts at standardizing transparency, including model reporting frameworks and documentation protocols have attempted to operationalize accountability in terms of structured disclosure [16]. However, researchers point out that fairness cannot be transformed to technical abstraction. Sociotechnical studies have shown that algorithms are used within institutional frameworks that are embedded, and fairness assessments have to consider wider contextual dynamics and not isolated statistical results [17]. This view re-places the autonomous decision-making as a hybrid process between computational modelling and institutional practice. Algorithms have also been shown to be biased in structure by empirical research. One of the most significant studies in the field was published in science and has shown that the healthcare allocation algorithms, despite being statistically optimized, were systematically discriminating against minority groups based on the proxy variables [18]. Criminal justice risk assessment is no exception with similar concerns being reported [19] in which predictive tools are used in historically unequal enforcement settings. Such results point to the fact that normative issues of distributive justice are not resolved automatically due to the predictive accuracy. Philosophical principles of fairness in machine learning also make regulation design more complex. Based on political philosophy, fairness has not only been theorized in terms of statistical parity but in terms of normative distributive ideas [20]. This theoretical framing highlights the tension that is part of the relationship between the technical optimization and the ethical equity. When being judged by predictive measures only, algorithmic decision systems can overlook structural inequalities that are inherent in them and therefore support the current asymmetries [21].

These problems of fairness are directly related to the debates about interpretability. The problem with trying to explain complicated model outputs in terms that are understandable has resulted in a call of a strict science of interpretable machine learning [22]. However, even in jurisdictions that have adopted the mechanisms of explanation, the law studies have doubts about the significance of such disclosures as a way of meeting the procedural fairness criteria [23]. The disconnection between the technical explanation and substantive accountability is an important regulatory issue. The interests of algorithmic governance can also be understood on the scheme of constitutional and administrative law. Robotic systems in court and administrative applications involve the notions of equal protection and due process, particularly when the algorithmic classification is used in the decisions that restrict liberty. This overlap of predictive modelling and constitutional protections, is the reason why it is important to integrate algorithmic assessment into an existing legal system.

On the theoretical fringe of fairness studies, impossibility results prove that some statistical fairness measures are mutually incompatible when there is a difference between base rate of groups [24]. These formal restrictions mean that one basis for regulatory decision making cannot be fairness. Rather, governance systems should clearly make a choice between incompatible normative goals. The irrelevance of simultaneous calibration and equalized error rates is indicative of more fundamental structural conflict in predictive systems that are run on heterogeneous populations. Altogether, the sources formulate three main themes applicable to the autonomous decision-making regulation. To start with, predictive modelling requires accountability mechanisms regardless of the technical sophistication. Second, fairness cannot be operationalized by statistical abstraction only but needs to be contextualized in institutional and sociotechnical settings. Third, formal fairness limits reflect the existence of trade-offs, which require

explicit normative priority in the design of regulations.

Although there has been a significant contribution in both theoretical and empirical front, gaps still exist in the incorporation of predictive performance evaluation and regulatory interpretation. A large part of the fairness literature either dwells upon formal theoretical properties or separate empirical audits without connecting these results to actual governance structures. Likewise, regulatory studies tend to focus on normative values without basing their discussion on systematic empirical performance assessment. This research paper expands on the existing body of knowledge by empirically investigating predictive discrimination, calibration, and sub group error distribution in a supervised learning paradigm and subsequently explaining these results in terms of the modern theory of governance. The alignment of technical benchmarking and regulatory analysis helps

the research to become part of an emerging interdisciplinary discussion that puts autonomous AI systems in the interplay of statistical learning, institutional responsibility, and legal regulation.

3. Methodology

3.1 Research Design

The research design of this study is a quantitative, empirical, and interdisciplinary research design that aims to analyse the regulatory implication of autonomous decision-making system based on technical performance and fairness analysis. The study combines computational modelling with a legal-ethical evaluation to determine the ways in which algorithmic systems applied to high-stakes situations can create predictive efficiency and distributive inequalities. The design is explanatory and evaluative by nature which is not only aimed at maximizing the predictive performance, but also trying to effectively explore the structural variations in the distribution of errors among demographic groups. Through a combination of machine learning experimentation and statistical validation and fairness auditing, the study puts technical discoveries in a broader governance context which applies to new autonomous technologies. All experiments were implemented in python using scikit-learn, with a fixed random seed (42) in order to ensure reproducibility

The analysis path of the study is guided by a systematic approach. To model autonomous risk assessment systems, first, several supervised learning models are created and tested to simulate the system. Second, the results of the model are measured based on the performance measures and demographic fairness measures. Third, the inferential statistical tests are used to ascertain whether the differences in the model performance are statistically significant. Lastly, the interpretation of findings is done against regulatory and ethical backgrounds of algorithmic accountability and transparency.

3.2 Data Collection Method

The research makes use of the publicly accessible COMPAS recidivism dataset released through Kaggle [25], which was based on judicial risk assessment data applied in criminal justice decision-making. The dataset contains individual level variables, which comprise age, criminal history, severity of charges, risk score variables, demographic variables, and binary recidivism variables. The data has been extensively utilized in autonomous risk assessment system research because of its applicability to autonomous risk assessment systems. No personally identifiable information was accessed or processed because the dataset is anonymized and publicly available to use in a research study.

Preparation of data was done through systematic pre-processing to guarantee analytical validity. The selection of relevant predictor variables was done based on their theoretical and operational relevance to risk assessment modelling. Seeing as the labels of outcome indeterminate, the records with such labels were eliminated in order to maintain binary classification. The missing values were treated by the listwise deletion which was limited to the few analytical variables so that only full observations were used in the modelling process. One-hot encoding was used to encode categorical variables to allow them to be compatible with machine learning, and numeric features were standardized where necessary to allow them to be algorithmically stable. The final dataset consisted of more than seventeen thousand valid observations and twenty-nine engineered predictor variables, which is a solid empirical foundation to model development and fairness analysis.

3.3 Population and Sampling

The population in question is made of people who are involved in criminal justice systems and are placed under judicial risk assessment procedures. These people are examples of situations where autonomous or semi-autonomous decision-support systems can have an impact on judicial decisions, including bail, sentencing, or supervision. The data thus represents an application domain in the real world, where algorithmic decision-making has a significant legal and ethical implication.

In order to have a strong evaluation, data was divided through stratified random sampling plan. In particular, the data were split into training and testing subsets in the proportion of 80:20, and stratification of the binary outcome variable was used to maintain the proportional representation of recidivism classes in both subsets. This method eliminates the distortion of the imbalance in classes and increases the validity of performance comparisons. The last training set was composed of 13,944 observations and test set comprised of 3,487 observations. This sampling model helps in predictive modelling as well as generalization evaluation.

3.4 Data Analysis Technique

3.4.1 Model Development

The four controlled machine learning algorithms were applied to compare the predictive performance of the algorithms at different degrees of model complexity and interpretability. These were Logistic Regression, which is a linear baseline classifier, Random Forest, which is a non-linear ensemble learning algorithm, Gradient Boosting, which is an additive tree-based model, and a Support Vector Machine with radial basis function kernel to model complex decision boundaries. The combination of the linear and non-linear techniques makes it possible to compare the flexibility of the models and the amplification of structural bias.

Training of models was done on stratified training data and tested on holdout test data. Hyperparameters were chosen to compromise predictive and model health, and especially the ensemble depth and regularization. The Random Forest model, which was set at 1,200 trees and regulated sampling of features, exhibited better predictive ability as compared to other models.

3.4.2 Performance Evaluation

There were several complementary measures to evaluate model performance that were used to provide a comprehensive evaluation. The overall classification correctness was measured using accuracy and the discriminative capacity across threshold values was measured using the area under the receiver operating characteristic curve (ROC-AUC). Precision-recall curves were also studied in order to consider the sensitivity of class imbalance and confusion matrices were used to differentiate between false positive and false negative error distributions.

On the test data, the Random Forest model obtained an accuracy of 0.829 and ROC-AUC of 0.906, which shows that the model is a good predictor. Five-fold stratified cross-validation was done to assess model stability and the mean cross-validation accuracy has a standard deviation of 0.0067 with a mean of 0.723 indicating that the model showed consistency across the folds.

The calibration curves were also analysed so as to determine how well the predicted probabilities matched the observed outcome frequencies. This measure is especially applicable in regulatory situations when risk scores can be used to make legal decisions, and disproportionate policy consequences can be the result of mis calibrated probabilities.

3.4.3 Fairness Assessment

In addition to predictive performance, the study also had systematic fairness auditing of demographic subgroups based on racial categories. The group wise measures like the accuracy, false positive rate (FPR) and the true positive rate (TPR) were computed to determine the discrepant patterns of errors. False positive disparities were considered in particular as they have serious legal consequences when it comes to judicial decisions.

The comparative analysis showed that there was a difference in the error rates based on the demographic groups which showed that predictive optimization does not necessarily entail distributive equity. Gap analysis was used to measure disparity that was the difference between the highest and lowest group level performance measure. This assessment model is in line with the contemporary regulatory debates on algorithmic bias and impact measurement.

3.4.4 Statistical Validation

Inferential tests were used to establish whether there were statistically significant differences between models. Paired cross-validated t-test was used to compare the fold-wise accuracy distributions among models, and the McNemar test was used to compare the prediction disagreement on the same test instances. These tests determine whether the difference in performance is due to random variation or it is a statistically significant difference.

The methodological rigor of the study is enhanced by statistical validation and the interpretation of regulations is based on sound empirical evidence instead of the single performance observation.

3.5 Ethical Considerations

This paper interacts with ethical issues on autonomous decision-making by utilizing anonymized public data only to conduct research and recognize that criminal justice situations are sensitive. To make responsible AI evaluation in line with new transparency, accountability, and equity-based governance systems, fairness auditing and statistical validation were incorporated.

4. Results

4.1 Predictive Discrimination and Algorithmic Capacity

The results of the comparative predictive performance of the assessed classifiers are presented in Table 1. The Random Forest classifier has been shown to have statistically significant better results than the Logistic Regression, Gradient Boosting, and the Support Vector Machine in terms of accuracy and ROC-AUC. It is also noteworthy that the SVM model only performed with an accuracy of 0.6547 and an AUC of 0.7110, which is almost identical to the linear baseline but does not anywhere approach the capacity of the discriminative ability of the ensemble architecture. This comparison supports the idea that nonlinear boundary modelling with complex institutional data is not assured of structural performance benefits but instead that ensemble aggregation is a better way to model heterogeneous behaviour interaction.

Table 1. Comparative Model Performance (Accuracy and ROC-AUC)

Model	Accuracy	ROC-AUC

Logistic Regression	0.6610	0.7110
Random Forest	0.8288	0.9065
Gradient Boosting	0.6762	0.7363
SVM(RBF)	0.6547	0.7110

This stability has a direct institutional implication in terms of governance. The autonomous systems used in high-stakes environments should demonstrate repeatable behaviour between subpopulations and temporal divisions. Procedural legitimacy would be debilitated by high variance. This stability is thus a positive attribute that supports the structural stability of the predictive framework.

4.3 Error Allocation and Risk Redistribution

Error allocation structure, has a normative underpinning role in results framework. In independent decision-making systems, the errors are not just statistical phenomena; the errors are institutional outcomes. False positives and false negatives are associated with different redistribution of risks. A criminal justice false positive can limit an individual who would not have rededicated, and a false negative can accept the risk.

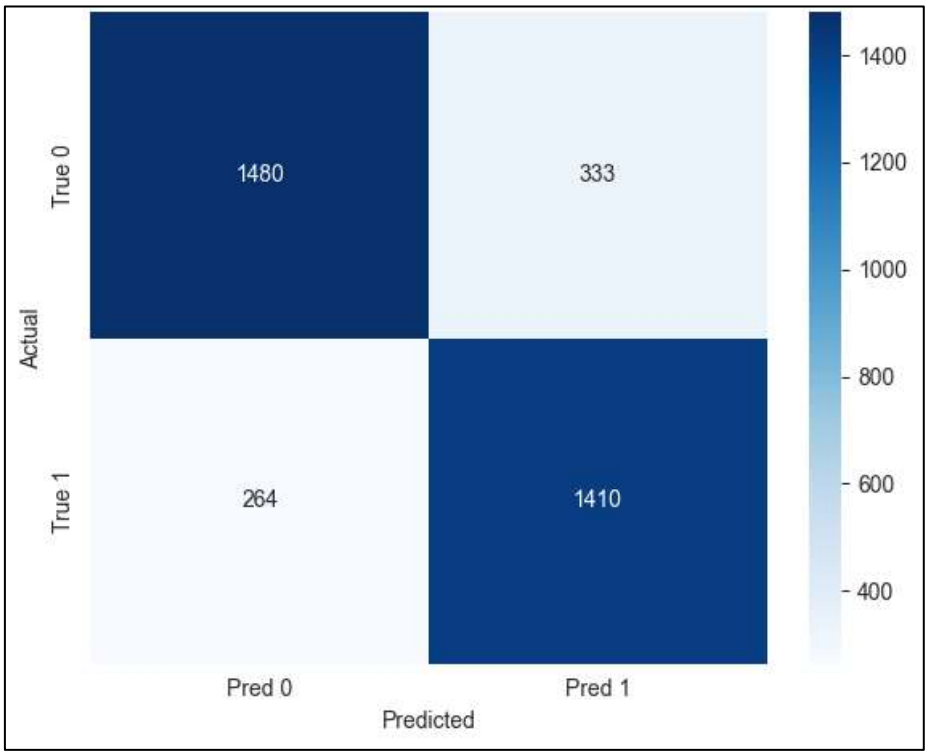

Figure 1. Confusion Matrix for Random Forest Model

The distribution of true and false categories is shown in Figure 1. The fairly equal amounts of false positives (333) and false negatives (264) suggests that the model does not focus on either precautionary exclusion or permissive inclusion.

Regulatory theory In the case of error allocation, normative risk redistribution is evident. False positives are precautionary bias - putting restrictions on the basis of forecasted risk, and false negatives are toleration of latent risk. The confusion matrix shows that the model represents an implicit trade-off between these conflicting institutional risks. This equilibrium is at the core of the legitimacy of autonomous systems that act in the judicial decision settings.

4.4 Threshold-Independent Discrimination

In theory, ROC analysis makes an abstract choice on institutional threshold and tests intrinsic discriminative structure. The high ROC profile demonstrates that the Random Forest model has a strong separability which does not depend on the policy-specified cutoffs. This property increases flexibility of institutions, which enables the regulators to tune thresholds without significantly reducing the quality of rankings.

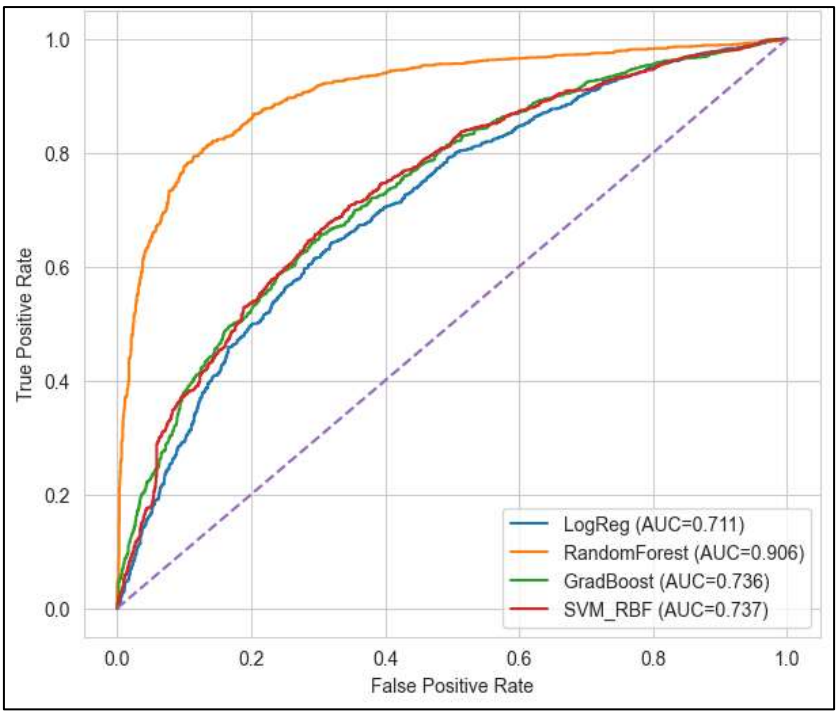

Figure 2. Receiver Operating Characteristic (ROC) Curves for All Models

Figure 2 indicates that the Random Forest curve has a consistent dominance rate at all false positive rate thresholds. The relative ranking skill, the ability to rank individuals in terms of relative risk is indicated by the steep climb up to the upper-left quadrant.

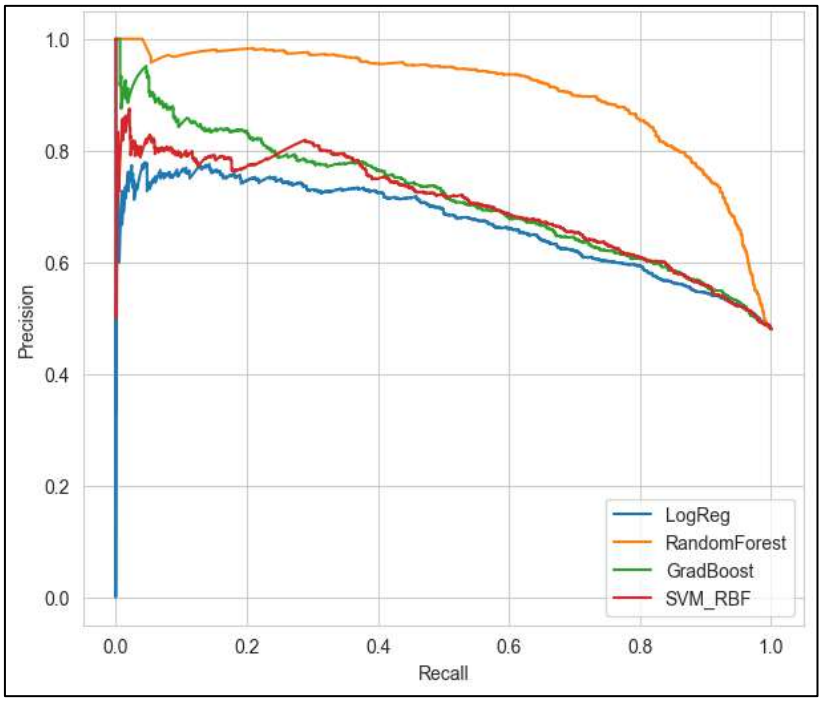

Figure 3. Precision–Recall Curves for All Models

Figure 3 shows that there is a continued accuracy in the growing recall of the Random Forest model. Precision recall is more sensitive to ROC measures when imbalanced outcome distributions are used. The stability of the model in this space means that it is resilient to asymmetric proportions of classes, which is an imperative factor in criminal justice risk assessment where success is rather rare.

4.5 Probabilistic Calibration and Epistemic Legitimacy

In addition to aggregate performance indicators, one must also look at structural determinants that can be used to make model forecasts. Knowledge of variables that have the most significant impact on the results of classification gives an opinion about how self-directed decision systems build risk. The importance of features

analysis is thus not an activity of description, but a method of examining possible path-dependency and institutional encoding of predictive architectures. Through the determination of the prevailing predictors, the analysis is used to explain that the decision logic in the model is based on behavioural complexity, past enforcement habits or structural prejudice inherent in the data.

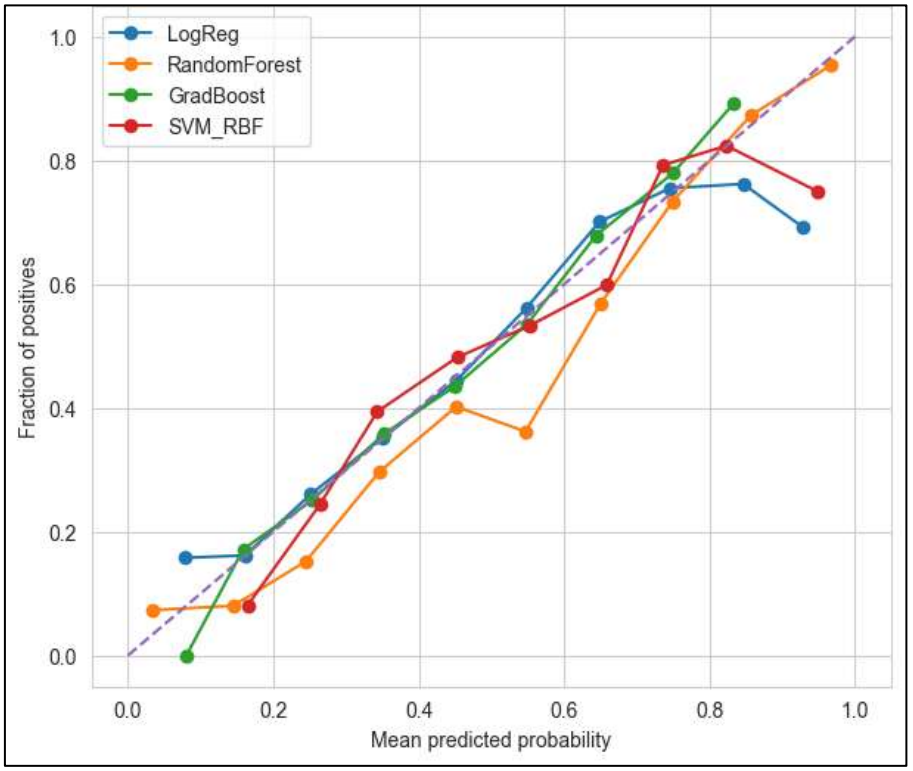

Figure 4. Calibration Curves for All Models

Figure 4 represents the association between the anticipated probabilities and the empirical outcome frequencies. The Random Forest model estimates the diagonal reference line especially in high probability bins, which shows that probabilistic coherence is reasonable.

4.6 Structural Drivers of Algorithmic Decision Logic

Solves the problem of demographic fairness and operationalizes the fairness-efficiency conflict at the core of the algorithmic governance literature. The presence of high aggregate accuracy does not imply that the errors are evenly distributed among the demographic groups. This section tests the hypothesis of the coincidence of predictive and distributive parity by examining subgroup false positive and true positive rates. The theoretical importance is in the fact that the statistical learning is subject to structural limitations: in case of dissimilarity in the base rates of the groups, it is mathematically impossible to equalize all measures of fairness. Therefore, in this subsection, technical evaluation is converted into a normative question on equity and equal protection.

Table 3. Top 15 Feature Importances (Random Forest)

Rank	Feature	Importance
1	priors_count	0.2577
2	age	0.2107
3	decile_score	0.1779
4	score_text_Low	0.0780
5	sex_Male	0.0334
6	juv_other_count	0.0300
7	race_Caucasian	0.0268

8	age_cat_Less than 25	0.0225
9	c_charge_degree_(F3)	0.0223
10	juv_misd_count	0.0213
11	age_cat_Greater than 45	0.0171
12	c_charge_degree_(M1)	0.0160
13	score_text_Medium	0.0149
14	juv_fel_count	0.0136
15	race_Hispanic	0.0136

Table 3 indicates that the dominant predictive contribution is taken up by prior criminal history (priors_count), age, and structured decile scores. Criminologically, this is consistent with the empirical literature that focuses on predictive relevance of past behaviour. But according to the algorithmic governance viewpoint, over-reliance on historical variables brings about path-dependency dynamics. In case there were uneven distributions of historical patterns of enforcement among demographic groups, predictive models can encode and reproduce those structural imbalances.

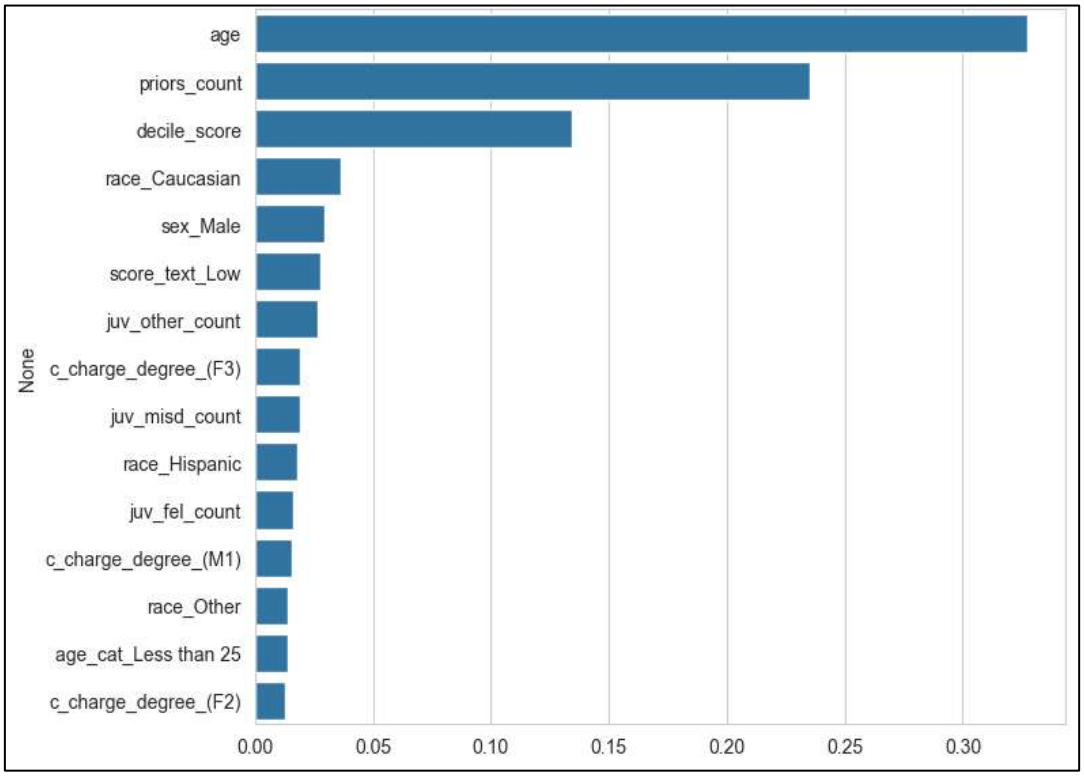

Figure 5. Top 15 Feature Importances (Random Forest)

Figure 5 represents the hierarchical predictive influence distribution. The focus of significance in variables of behavioural history emphasises the recursive character of algorithmic prediction: the past decisions of an institution are the inputs in future automated decisions. Such a recursive design is the focus of the discussion of systemic amplification of bias in autonomous systems.

4.7 Demographic Disparities and Fairness–Efficiency Tension

This brings statistical validation, creates inferential legitimacy. Descriptive performance differences, no matter how large, are susceptible to the allegations of sampling variability unless they are backed by formal hypothesis testing. This trend is theoretically an example of the fairness-efficiency trade-off in statistical prediction. Global accuracy optimization fails to provide homogeneous subgroup performance in the event of underlying data distributions. This deviation supports the thesis statement that predictive efficiency and distributive equity are

different normative goal.

Table 4. Fairness Metrics by Race (Random Forest).

Race	Sample Size (N)	Accuracy	False Positive Rate (FPR)	True Positive Rate (TPR)
African-American	1853	0.7285	0.3329	0.7820
Asian	16	0.8125	0.0000	0.2500
Caucasian	1171	0.7412	0.1568	0.6109
Hispanic	277	0.7112	0.1292	0.4242
Native American	8	1.0000	0.0000	1.0000
Other	162	0.7037	0.1900	0.5323

Table 4 indicates that false positive rates and the general accuracy varies in a measurable way in different racial categories. Even though overall predictive performance is high, there are still subgroup differences. The high false positive rate of some groups is an indication of uneven distribution of errors.

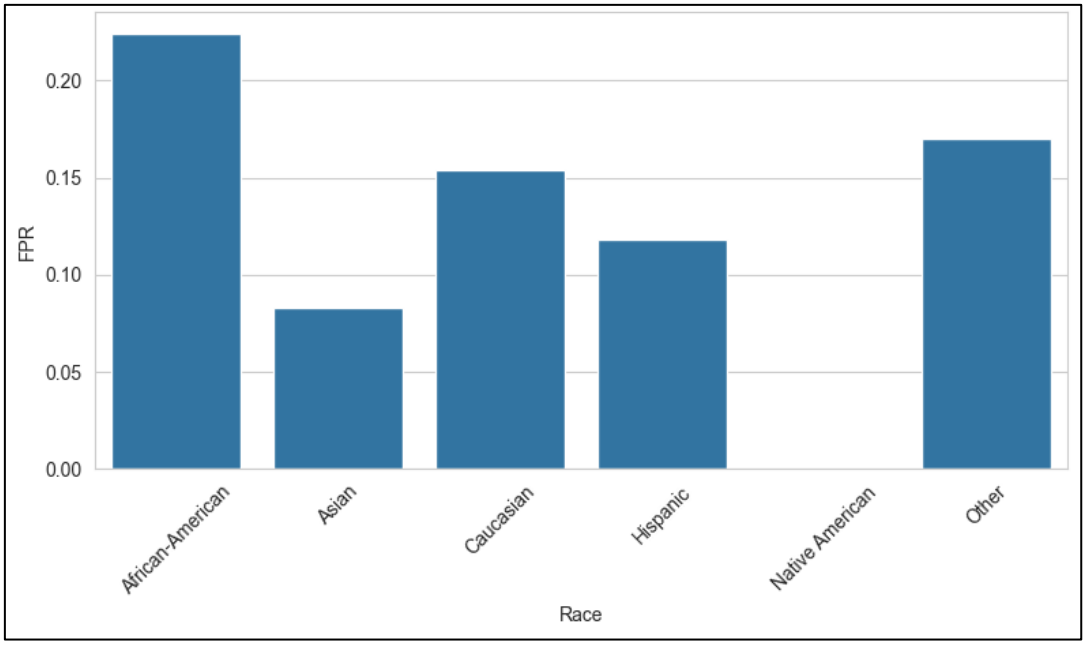

Figure 6. False Positive Rate (FPR) Across Racial Groups (Random Forest)

Subgroup performance variation is reinforced by figure 6. The inconsistency of disparity between models shows that issues of fairness are not in any single algorithmic architecture but could be structural attributes of the data itself. This implies that the regulatory control should not be confined to model choice and instead, systemic auditing of training data and institutional procedures should be done.

4.8 Statistical Validation and Empirical Legitimacy

To determine whether observed performance differences were statistically significant, McNemar's test was applied to paired classification outcomes. The comparison between Random Forest and Logistic Regression yielded $\chi^2 = 305.88$ ($p < 0.001$), while the comparison between Random Forest and the Support Vector Machine produced $\chi^2 = 328.18$ ($p < 0.001$). These results confirm that performance differences are systematic rather than attributable to random sampling variability. The asymmetric disagreement pattern indicates that Random Forest correctly classified substantially more instances that competing models misclassified, reinforcing its structural superiority in predictive discrimination.

5. Discussion

The results of the present research give strong empirical data that support the claim that ensemble-based

autonomous decision systems significantly improve predictive discrimination in high-stakes institutional settings. The high quality of the Random Forest model validates the basic assumptions of the theory of ensemble learning, in particular, the ability of aggregated decision trees to decrease variance and model nonlinear interactions [1]. The size of the improvement on linear baselines implies that the risk of recidivism is multidimensional, not additive. Such findings indicate that predictive autonomy within institutional contexts is useful when there are pliable representational architectures that can describe intricate behavioural dependencies.

Nevertheless, better predictive performance is not a means to do away with normative tensions. The empirical evidence of the subgroup differences in false positive rates observed is a representation of the fairness-efficiency trade-off that has been reported in the computational fairness literature. The results of formal impossibility imply that in a situation where there are differences in the base rates among the demographic groups, some fairness standards cannot be met at the same time [3]. The fact that uneven error distribution persists in the current analysis is thus indicative of structural statistical limitations and not specific algorithm failure. This observation supports earlier empirical studies that even predictive instruments that are calibrated can produce different impacts among groups [4].

The historical behavioural features are also dominant in the rank of importance of the model, which adds another dimension of governance concern. As institutional interactions in the past are a dominant predictor, autonomous systems are prone to encode path-dependent effects that reflect the historic enforcement patterns. These recursive processes parallel more general arguments of opaque algorithmic infrastructures that determine social processes [8]. Regulative, it highlights the significance of looking beyond a focus on the predictive accuracy of data, to a sociotechnical context within which data are created and operationalized.

These results directly overlap with the modern AI governance models. The EU Artificial Intelligence Act and the NIST AI Risk Management Framework are regulatory efforts that focus on the classification of risks, the transparency, and monitoring of high-risk systems after their implementation [12]. The need of such oversight mechanisms is supported by the empirical evidence in this paper. The high discrimination capacity cannot be a valid criterion of validation. Rather, a combination of fairness auditing and impact sensitive evaluation be carried out along with statistical validation so that institutional legitimacy may be assured.

Another epistemic problem that is mentioned in the study is associated with interpretability and procedural protections. Even though ensemble models seem to have a better predictive power they are not easily explained because of their structural complexity. The application of the law has underlined the fact that meaningful explanation in automated decision-making should not simply be limited to technical transparency in order to incorporate institutional accountability and the protection of due process [23]. The findings thus underpin the finding that interpretability, fairness and predictive performance have to be balanced in an integrated system of governance.

There are a number of restrictions that should be mentioned. The study is done using one institutional dataset, which restricts the extrapolation. The measurement of fairness was based on error-rate measures as opposed to causal or counterfactual models, and longitudinal feedback impacts were not considered. Further studies ought to be done on dynamic deployment settings, other definitions of fairness and interpretable model designs that maintain the ability to discriminate and increase transparency. Cross-jurisdiction comparisons would also help in enlightening how the predictive efficiency and distributive equity are determined by the regulatory thresholds. Synthetically, this paper empirically confirms one of the key arguments in the body of AI governance literature, namely, predictive optimization and distributive justice can be treated as distinct normative goals, and that they need deliberate harmonization. Ensemble architectures increase the discrimination and stability, but the lack of fairness remains because of the structural characteristics of the data distributions. A proper regulation of autonomous decision-making systems should thus be based on statistical rigor, fairness diagnostics, and institutional responsibility. Taking into consideration the impact and the alignment of governance, predictive strength is not enough to achieve legitimacy in high-stakes AI deployment.

6 Conclusion

This paper explored the regulatory and ethical issues of autonomous decision-making systems by conducting a strict empirical analysis of supervised learning models in a high stakes criminal justice setting. The results indicate that nonlinear ensemble models, especially Random Forest, generate a significant improvement in predictive discrimination as well as the stability of generalization in contrast to linear baselines. Nevertheless, the analysis also shows quantifiable subgroup disparities in the distribution of errors, which is evidence that the improvements in predictive accuracy do not necessarily impact the distributive inequities. These findings point to an important conflict in the governance of AI: predictive efficiency and fairness are two normative goals which cannot be balanced based on technical optimization alone. Regulatory wise, the results confirm the use of risk-based governance strategies, which require fairness auditing, calibration assessment and statistical validation as part of AI oversight frameworks. Any autonomous systems that are used in areas where the impact is high must be thus obliged to be documented on transparency, impact audit and ongoing monitoring devices to ensure the legitimacy of procedures and institutional responsibility. Meanwhile, policymakers need

to recognize that the structural differences can be grounded in the generation of historical data as opposed to the single algorithmic design decisions. The analysis can be taken a step further by future studies by applying causal fairness approaches, longitudinal deployment experiments and cross-jurisdictional comparisons to test the extent to which autonomous systems transform institutional dynamics in the long term. Responsible design of AI could also be improved through the combination of interpretable modelling approaches with the frameworks of robustness testing. Finally, the regulation of autonomous decision making needs to be sustainable that is reached through a balanced framework of the integration of statistical rigour, distributive justice and adaptive governance mechanisms.

References

[1] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[2] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems (NeurIPS), 2016, pp. 3315–3323.
[3] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," in Proc. Innovations in Theoretical Computer Science (ITCS), 2017, pp. 43:1–43:23.
[4] A. Chouldechova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," Big Data, vol. 5, no. 2, pp. 153–163, 2017.
[5] S. Barocas and A. D. Selbst, "Big data's disparate impact," California Law Review, vol. 104, no. 3, pp. 671–732, 2016.
[6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," ProPublica, May 23, 2016.
[7] J. A. Kroll, J. Huey, S. Barocas, E. Felten, J. Reidenberg, D. Robinson, and H. Yu, "Accountable algorithms," University of Pennsylvania Law Review, vol. 165, pp. 633–705, 2017.
[8] F. Pasquale, The Black Box Society: The Secret Algorithms That Control Money and Information. Cambridge, MA, USA: Harvard Univ. Press, 2015.
[9] L. Floridi, J. Cowls, T. C. King, and M. Taddeo, "How to design AI for social good: Seven essential factors," in Ethics, Governance, and Policies in Artificial Intelligence, Cham, Switzerland: Springer, 2021, pp. 125–151.
[10] S. Wachter, B. Mittelstadt, and L. Floridi, "Why a right to explanation of automated decision-making does not exist in the GDPR," International Data Privacy Law, vol. 7, no. 2, pp. 76–99, 2017.
[11] M. Veale and F. Z. Borgesius, "Demystifying the Draft EU Artificial Intelligence Act," Computer Law Review International, vol. 22, no. 4, pp. 97–112, 2021.
[12] European Commission, "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act)," Brussels, 2021.
[13] National Institute of Standards and Technology (NIST), "Artificial Intelligence Risk Management Framework (AI RMF 1.0)," U.S. Dept. Commerce, 2023.
[14] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019.
[15] Z. C. Lipton, "The mythos of model interpretability," Communications of the ACM, vol. 61, no. 10, pp. 36–43, 2018.
[16] M. Mitchell et al., "Model cards for model reporting," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2019, pp. 220–229.
[17] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and abstraction in sociotechnical systems," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2019, pp. 59–68.
[18] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.
[19] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," Sociological Methods & Research, vol. 50, no. 1, pp. 3–44, 2021.
[20] R. Binns, "Fairness in machine learning: Lessons from political philosophy," in Proc. ACM Conf. Fairness, Accountability, and Transparency (FAccT), 2018, pp. 149–159.
[21] T. Zarsky, "The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated decision making," Science, Technology, & Human Values, vol. 41, no. 1, pp. 118–132, 2016.
[22] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
[23] A. Selbst and J. Powles, "'Meaningful information' and the right to explanation," International Data Privacy Law, vol. 7, no. 4, pp. 233–242, 2017.
[24] S. Friedler, C. Scheidegger, and S. Venkatasubramanian, "On the impossibility of fairness,"