Identifying biases in legal data: An algorithmic fairness perspective

JACKSON SARGENT[*], University of Michigan

MELANIE WEBER[†], Princeton University & Claudius Legal Intelligence

The need to address *representation biases* and *sentencing disparities* in legal case data has long been recognized. Here, we study the problem of identifying and measuring biases in large-scale legal case data from an algorithmic fairness perspective. Our approach utilizes two regression models: A baseline that represents the decisions of a "typical" judge as given by the data and a "fair" judge that applies one of three fairness concepts. Comparing the decisions of the "typical" judge and the "fair" judge allows for quantifying biases across demographic groups, as we demonstrate in four case studies on criminal data from Cook County (Illinois).

Additional Key Words and Phrases: Algorithmic fairness, Legal data, Representation bias, Sentencing disparities, Data transparency

1 INTRODUCTION

A cornerstone of the development of trustworthy and discrimination-free machine learning and artificial intelligence is the consideration of biases in the training data. As artificial intelligence enters the legal space, it is essential to recognize biases in legal data and ensure that they are not replicated and reinforced with legal technology [7, 13, 18]. Furthermore, understanding biases in legal data and developing discrimination-free technology could help the legal space to become fairer and more widely accessible.

We typically find two types of biases in legal data: First, *representation biases*, i.e., certain social groups are over- or underrepresented in a data set. Second, *sentencing disparities*, i.e., the outcome of legal proceedings for similar cases varies across social groups. Representation biases may reflect disparities in policing (arrest rates) or in offense rates. While potential differences in offense rates across demographic groups (e.g., male and female for specific crime types) are difficult to evaluate, policing disparities have been studied with data-driven methods through the analysis of arrest data. The need to address sentencing disparities has long been recognized. In the context of federal sentencing, the Sentencing Reform Act was passed in 1984 as part of the Comprehensive Crime Control Act with the aim of mitigating sentencing disparities. This has resulted in the establishment of the United States Sentencing Commission in 1987 to publicize federal sentencing guidelines. Today, both representation biases and sentencing disparities are active areas of legal scholarship [2, 3, 9, 16, 19, 21, 22].

Recently the data-driven analysis of representation biases and sentencing disparities has gained a surge of interest [4, 7, 21]. The present study is guided by the following question:

> *How can we efficiently identify and measure biases in large-scale legal data?*

Much of the existing body of literature focuses on the quantitative analysis of sentencing disparities [9, 19, 21], for instance, by analyzing correlations between the characteristics of the defendant – and sometimes the judge – with the case outcome, e.g., the length of the sentence. Here, we take a different approach. We propose to compare the decisions of a "fair" judge with the judicial decisions in our data set. For this, we analyze two models: A baseline model trained via an unconstrained regression approach on the data ("average" or "typical" judge) and a second model trained with *fairness constraints* ("fair" judge).

[*] Work done while intern at Claudius Legal Intelligence.

[†] Corresponding author.

In addition to studying sentencing disparities via algorithmic fairness metrics, we also analyze representation bias by comparing the performance of the baseline model with a second, balanced model. Finally, we compare with a balanced classifier that was trained with fairness constraints – combining both bias mitigating approaches.

In contrast to much of the prior literature, which analyzes sentencing disparities mostly with respect to the case outcome, we focus on other legal proceedings. In particular, we analyze, (1) whether bond without upfront payment was granted and (2) whether, upon appeal, the charge was reduced. The choice to focus on legal proceedings with binary outcomes was motivated by their simplicity and suitability for existing fairness metrics and methodology, which typically assumes categorical, rather than continuous outputs (such as, e.g., the sentence length).

The study's goal is explicitly not the predictive modeling of the judicial system, nor does it provide a final solution for debiasing legal data. Instead, we focus on developing protocols and pipelines for identifying and understanding biases in legal data, mainly focusing on *data transparency* and *awareness of existing biases*.

1.1 Related Work

The data-driven analysis of biases in legal data has recently gained a surge of interest. Notable work that quantitatively analyses sentencing disparities includes [8–10, 19, 21]. In [19, 21], the focus lies mainly on how the characteristics of the defendant influence the likelihood of certain case outcomes. In contrast, [8–10] evaluate the characteristics of both the defendant and the judge. Notably, [8, 10] consider identifying information on judges. The recently initiated *JUSTFAIR* project [8] collects large-scale criminal data. It includes, besides sentencing information, also demographic information and criminal history for the defendant, as well as demographic information on the judge.

The proposed approach applies algorithmic fairness tools, which were developed in several seminal works, notably [11, 14]. [6] studies the problem of mitigating biases in training data in the broader machine learning context. [17] broadly discusses the idea of detecting discrimination and bias with algorithmic tools. The Fairlearn toolkit [5] implements a range of fairness metrics.

1.2 Summary of contributions

The goal of this study is to develop pipelines for efficiently identifying and measuring biases in large-scale legal data. Specifically, our contributions are as follows:

(1) In four case studies, we compare the decisions of a "typical" judge with those of a "fair" judge that applies one of three concepts of fairness. This setup allows for quantifying biases in the underlying criminal case data and gives insight on the suitability of these notions for legal data.
(2) We analyze the impact of representation biases on fairness in legal cases by incorporating data balancing into our mitigated ("fair") models.

To the best of our knowledge, the present study is the first to evaluate biases in legal data from an algorithmic fairness perspective.

2 METHODS

In this section we introduce the fairness concepts used in our study, as well as the methodology underlying both the baseline and the "fair" classifiers. We adapt an experimental approach introduced in [1] to the training of fair classifiers on legal data. All models were implemented in Python.

2.1 Notions of fairness

We begin by formally introducing the fairness concepts considered in this study. Since the study focuses on fairness aspects of judicial decisions across demographic groups, we focus on *group fairness*, as opposed to individual fairness notions. Below, we give formal definitions for a prediction task with binary outcomes ($\hat{Y} \in \{0, 1\}$) with respect to a binary protected characteristic ($A \in \{0, 1\}$).

The first notion we consider is *demographic parity*, which requires that the outcome is independent of the protected characteristic, i.e.,

$$\mathbb{P}\left[\hat{Y} = 1 | A = 0\right] = \mathbb{P}\left[\hat{Y} = 1 | A = 1\right] \ . \tag{1}$$

Unfortunately, this notion is known to have serious shortcomings. In particular, [11] argues that demographic parity is not sufficient to ensure fairness (especially in the presence of representation biases), and may not allow for learning the optimal predictor. To mitigate these shortcomings, [14] introduces two additional fairness metrics: Equalized odds and equalized opportunity. *Equalized odds* equalizes both true positive and false positive rates across demographic groups; formally, we require

$$\mathbb{P}\left[\hat{Y} = 1 | A = 0, Y = y\right] = \mathbb{P}\left[\hat{Y} = 1 | A = 1, Y = y\right] \quad \forall y \in \{0, 1\} \ . \tag{2}$$

Equal opportunity is a relaxation of equalized odds, where only the true positive rates are equalized across demographic groups, i.e.,

$$\mathbb{P}\left[\hat{Y} = 1 | A = 0, Y = 1\right] = \mathbb{P}\left[\hat{Y} = 1 | A = 1, Y = 1\right] \ . \tag{3}$$

2.2 Baseline classifier

We consider two baseline classifiers: A *weighted logistic regression model* and a *gradient boosted decision tree*. Below, we briefly describe both approaches. For a more detailed overview on both approaches, see [15]. Our experiments use an implementation from the SCIKIT-LEARN package[1].

The first approach utilizes a *classical regression model* with a logistic loss function. The model returns, for a given case input, the probability of two binary outcomes: In the first experiment, whether bail without upfront payment was granted ("free bond") and in the second, whether the charge was reduced upon appeal ("charge reduction"). Here, we use a *weighted* approach, meaning that we incorporate weights in the training pipeline that are chosen inversely proportional to class frequencies in the input data. The class frequencies and weights are given in Appendix C.2.

Formally, the output is computed as follows: Let $x \in \mathbb{R}^d$ denote an input feature vector (where d denotes the number of features in the data, here $d = 7$); a and b the model's coefficients and P the output probability.

$$P(x) = \frac{e^{a+bx}}{1 + e^{a+bx}} \ . \tag{4}$$

Due to its simplicity, the logistic regression model is a suitable benchmark for basic classification tasks. The hyperparameters for the model are specified in Appendix C.1.

The second approach, a *gradient boosted decision tree model*, learns an ensemble of simple decision trees (decision "stumps"), which predict an outcome based on only one of the input features. The model combines these decision "stumps" iteratively into one predictor, which can then perform well on more complex classification tasks. In iteration j,

[1] https://scikit-learn.org/stable/

an update of the form

$$F_j(x) = F_{j-1}(x) + \gamma_j h_j(x) \tag{5}$$

$$\gamma_j = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, F_{j-1}(x_i) + \gamma h_j(x_i)) \tag{6}$$

is performed, where $h_j(x)$ is a decision tree, γ_j is a coefficient chosen to minimize a loss function L (in our experiments, the logistic loss) over a training set $\{(x_i, y_i)\}_{i=1}^{n}$. The gradient boosted decision tree approach implements a more complex model than the regression approach and is expected to have superior performance. We use the default hyperparameters for this model, as specified in Appendix C.1.

2.3 Classifier with fairness constraints

A first (naive) idea for training a "fair" classifier might be to simply ignore all protected characteristics, such as race or gender, when aiming to train an unbiased classifier. However, this approach does not mitigate biases effectively, since protected characteristics can typically be inferred from unprotected ones [20]. Instead, we analyze data biases using the classical algorithmic fairness notions introduced in section 2.1: *Demographic Parity*, *Equal Odds* and *Equal Opportunity*. For each legal setting and crime type, we train, in addition to the baseline, three "fair" classifiers using the three fairness metrics. We constrain each trained classifier to adhere to one of the fairness notions and evaluate the impact of the mitigating constraint with respect to all three fairness notions. In order to integrate the fairness constraints into our training pipeline, we follow the setup in [1] (*exponentiated-gradient reduction*[2]). We again consider both logistic regression and boosted decision trees.

2.4 Data balancing

Finally, we aim to account for *representation bias*, too. We balance data by *random subsampling*: To achieve an equal split between the two (binary) outcomes, we sample an appropriately sized subset from the data points with the more frequent outcome. We use NUMPY's *random shuffle*[3] function to perform the subsampling.

2.5 Evaluation metrics

We report results with respect to the following fairness metrics:

- Demographic parity difference: $\Delta_{DP} := \left|\mathbb{P}\left[\hat{Y} = 1|A = 0\right] - \mathbb{P}\left[\hat{Y} = 1|A = 1\right]\right|$
- True positive difference: $\Delta_{TP} := \left|\mathbb{P}\left[\hat{Y} = 1|A = 0, Y = 1\right] - \mathbb{P}\left[\hat{Y} = 1|A = 1, Y = 1\right]\right|$
- False positive difference: $\Delta_{FP} := \left|\mathbb{P}\left[\hat{Y} = 1|A = 0, Y = 0\right] - \mathbb{P}\left[\hat{Y} = 1|A = 1, Y = 0\right]\right|$

Naturally, a classifier with Demographic parity constraint is designed to mitigate Demographic parity difference; a classifier with Equal opportunity constraint to mitigate true positive difference and a classifier with Equalized odds constraint both true and false positive difference.

[2] https://github.com/fairlearn/fairlearn

[3] https://numpy.org/doc/stable/reference/random/generated/numpy.random.shuffle.html

3 LEGAL DATA

3.1 Data sources

	Total	Narcotics	Theft
Initiations	975,836	208,434	48,885
Dispositions	831,582	167,735	42,129

Table 1. **Size of raw data.**

	Charge reduction	Bond
Narcotics	34,806	35,002
Theft	12,862	12,363

Table 2. **Size of preprocessed data sets.**

The data used in this study is provided by the Cook County State's Attorney's Office (Illinois) and publicly available at the following websites:

(1) **Initiation:** https://datacatalog.cookcountyil.gov/Courts/Initiation/7mck-ehwz (download on March 30, 2021 at 6:00 PM CST)
(2) **Sentencing:** https://datacatalog.cookcountyil.gov/Courts/Sentencing/tg8v-tm6u (downloaded on March 30, 2021 at 6:00 PM CST)

Both data sets were created in February 2018. The first data set (*Initiations*[4]) contains 38 features, the second (*Dispositions*[5]) 33. We focus our analysis on seven features that are common between both data sets. The choice of features is described in detail in the following section. Furthermore, we consider only narcotics and theft crimes in our analysis. The selection of crime types was guided by the statistical power of the respective data subsets. Table 1 lists the size of the raw data. We use a 75%/ 25% training/testing split in the subsequent analysis.

3.2 Preprocessing

To generate the data sets used in the study, we filtered in terms of crime types and case outcomes: For crime types, we selected narcotics and theft cases. Furthermore, we generated data sets with binary case outcomes, with respect to charge reduction and bail. Table 2 lists the size of the data sets after preprocessing.

For the feature selection, we initially dropped unneeded features, such as dates, the primary charge flag (TRUE for all cases), the charge count, charge disposition (guilty plea for all cases), and redundant IDs. We then performed a correlation analysis (see Figure 5) of the remaining features and selected a subset with low pairwise correlations. Finally, we removed some erroneous data base entries with defendants of ages greater than 100 years old. More details on feature selection can be found in Appendix A.

We select cases for our target crime types based on the reported updated offense category, where we combined theft and retail theft into one category. We generated the *charge reduction* outcome by first selecting primary charges with plea bargains, which were accepted to receive reduced charges. We included cases for which the class of the charge was within a set of 5 degrees of severeness, where 0 is the most severe and 5 the least severe crime. By comparing the class of the charge in *Initiations* with that in *Dispositions*, we say that the charge was reduced, whenever the dispositions class was greater, or less severe, than the initiations class. For *bail*, we consider four types of bonds: I Bond, C Bond, D Bond, and "no bond". No bond is the most severe, where the defendant is denied any pretrial release. I Bonds are the least severe, allowing the defendant to be released without monetary bail. C Bonds and D Bonds are similar, with both

[4] An initiation is the decision to prosecute a defendant after a felony review by the State's Attorneys Office following an arrest; an indiction by a grand jury; or a direct filing by law enforcement (narcotics cases only).
[5] A disposition is the completion of the fact-finding process that leads to the resolution of a case.

requiring payment, where C Bonds require a full payment and D Bonds require only a partial payment. We consider binary outcomes of "no/ paid bonds" and "free bond". Here we combined C Bonds, D Bonds and "no bond" into the "no/ paid bonds" outcome and call I Bonds "free bonds".

Our study considers four demographic groups with respect to race and gender. In order to directly apply the classical fairness constraints introduced in section 2.1, we restrict ourselves to binary protected characteristics: We consider only cases in which the defendant identifies as female or male and as either white or black. For the former, the raw data reports the gender of defendants as either female or male. We excluded cases with gender reported to be unknown. For the latter, we focus on the two groups with the highest statistical power in the data set. The restriction to white and black defendants also follows a pattern in related legal scholarship, see, e.g., [12, 21], allowing for a placement of our study in the context of this body of literature. However, we note that a future, more comprehensive study should be more inclusive with respect to race and gender.

4 RESULTS

4.1 Statistical Analysis

We start with a preliminary statistical analysis of our legal data and compare the results across the two different crime types and the four demographic groups. We also analyze arrest numbers and the age of the defendant at incident.

We can see in Figure 1 that disparities are already evident in the arrest numbers across demographic groups: For both crime types, we observe that the number of arrests for black males are about three time higher than those of the other three groups. Comparing across races, we notice a larger number of black arrests; comparing across genders, we notice a larger number of male arrests. We further analyzed the number of arrests across incident cities in Cook county. The results can be found in Appendix B.

Next, we analyze the legal decisions in our case studies (*charge reduction* and *free bond*, for narcotics and theft crimes) across demographic groups. We observe that 48% of black men have their charge reduced, compared to 51% of white men, 49% of black women, and 52% of white women (see also Figure 3). The bail data shows that free bonds are given more often for theft crimes (68%), compared to narcotics crime (58%). Furthermore, black defendants are more likely to have restrictive bonds than other demographic groups, with 62% of black men receiving a free bond, compared to 64% of white men; and 49% of black women, compared to 53% of white women (Figure 4). We note that the overrepresentation of men (especially black men) in the data sets may significantly influence the observed trends for charge reduction and free bonds.

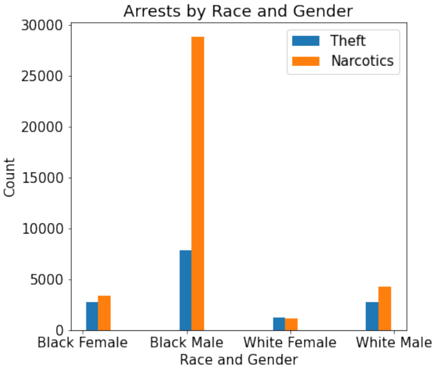

Fig. 1. **Arrest counts by race and gender for both crime types**, with Theft in blue and Narcotics in Orange. The figure shows that while arrest for other demographics are relatively close between crime types, about 3 times as many black Men are arrested for Narcotics crimes.

Other interesting trends can be found by looking at the reported age of those arrested. Figure 2 shows that those arrested for narcotics crimes are mostly young people between 20 and 30 years, across all demographic groups. For theft the trend is different: White people arrested for theft crimes are mostly around 30 years old, while the age at arrest for black people peaks at around 20 years old and again around 45 years.

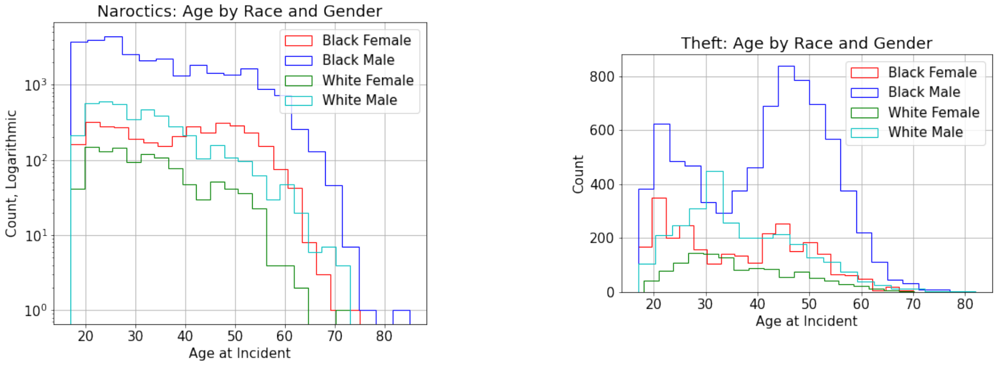

Fig. 2. **Age at arrest for narcotics (left) and theft (right).** The Narcotics graph is scaled logarithmic due to the much larger proportion of black Men present in the data. Narcotics crimes are mostly young people, while Theft appears to have two peaks among arrested blacks at around 20 and 45 years old. For arrested whites, the frequency peaks around 30 years old and decreases with age.

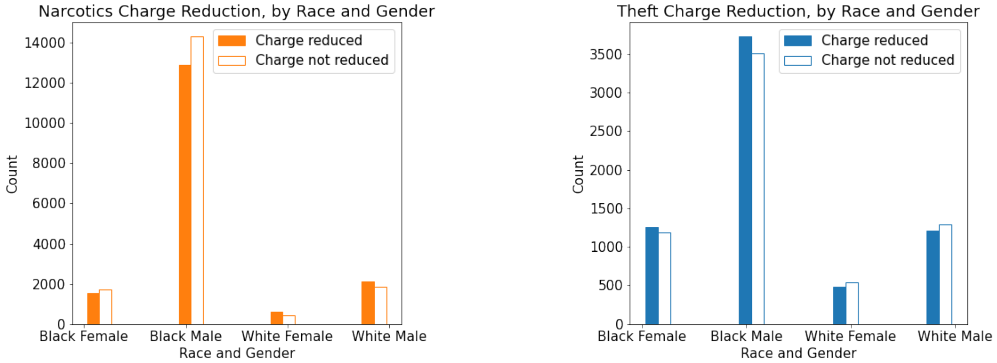

Fig. 3. **Binary outcomes for charge reduction.** Having a charge maintained means that the charge was not reduced. The percentage of arrested persons that had their charge reduced was comparable across demographic groups, except for black men in narcotics cases, who are more likely to *not* have their charge reduced.

4.2 Algorithmic analysis of data biases

After running our regression models on all four case studies (charge reduction/ free bonds and narcotics/ theft crimes), we found that on average, applying fairness constraints does succeed in mitigating biases. In addition, data balancing helps to mitigate representation biases and, on average, improves fairness metrics. We observe that our baseline classifiers, which represent the actions of a "typical" judge, perform the least fair. Both regression models (weighted logistic regression (Tab. 3, 5) and gradient boosted tree search (Tab. 4, 6)) achieve similar fairness results and both seem to be equally affected by balancing and fairness mitigation. Notably, data balancing improved, on average, the fairness metrics for both approaches (Tab. 3-6, first column).

As expected, applying a specific fairness constraint improves its target metric. We observe that applying a demographic parity constraint decreases the demographic parity difference, which indicates that our baseline was not classifying

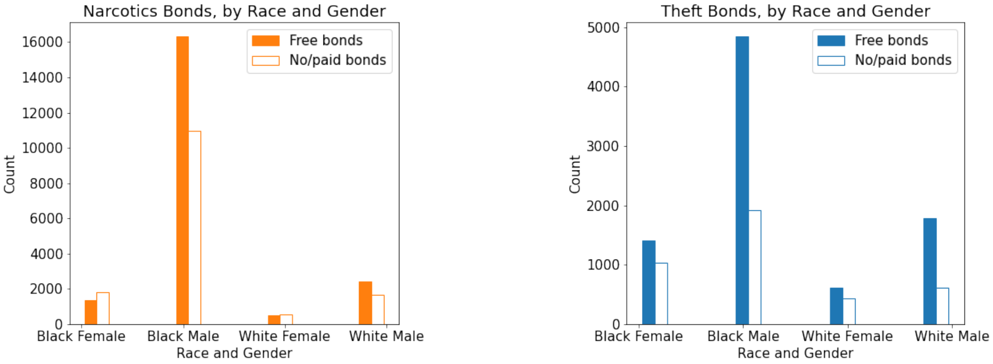

Fig. 4. **Bail outcomes for Narcotics and Theft crimes.** Free bonds refer to bail outcomes that grant freedom with no payment, and no/paid bonds refer to bail outcomes that either give no freedom or require a bond payment. Generally speaking, arrested persons are more likely to be granted free bonds in cases involving theft crimes, compared to cases involving narcotics crimes. Black persons are more often given restrictive bonds, compared to white persons.

	Baseline		Mitigated (DP)		Mitigated (EOdd)		Mitigated (EOpp)	
	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced
Narcotics								
Δ_{DP}	0.068	0.019	0.002	0.017	0.005	0.021	0.036	0.025
Δ_{TP}	0.014	0.019	0.091	0.022	0.068	0.009	0.045	0.010
Δ_{FP}	0.120	0.012	0.059	0.013	0.051	0.007	0.086	0.014
Theft								
Δ_{DP}	0.121	0.041	0.031	0.025	0.0448	0.006	0.057	0.045
Δ_{TP}	0.124	0.034	0.039	0.020	0.060	0.021	0.070	0.025
Δ_{FP}	0.108	0.100	0.013	0.058	0.019	0.021	0.034	0.100

Table 3. **Logistic Regression:** Results for predictive analysis (**Charge reduction**) for baseline in comparison with Demographic parity (DP), Equalized odds (EOdd) and Equalized Opportunity (EOpp) mitigations. Here, Δ_{DP} denotes Demographic parity difference, Δ_{TP} the True positive difference and Δ_{FP} the False positive Difference.

independent of race and gender (Tab. 3-6, second column). We also see that equalized odds mitigation decreases both true positive and the false positive difference (Tab. 3-6, third column), and that equalized opportunity decreases the true positive difference, meaning that the accuracy of our classifier varied across demographic groups (Tab. 3-6, fourth column). We further observe that equalized opportunity mitigation often decreases the true positive difference more than equalized odds, while either maintaining or increasing the false positive difference. This suggests that optimizing one fairness metric might worsen others. Alongside our mitigation techniques, we used data balancing to counteract representation bias. When applied to the baseline classifier, data balancing generally improves the fairness metrics.

Overall, *equalized odds with balancing* provided the best mitigation in our case studies. This suggests that our current judicial model could benefit from sentencing guidelines that equalize the rate of positive and negative outcomes across demographic groups. One of the primary objectives of sentencing guidelines is to reduce the sanctioning of innocent defendants, i.e., to reduce false positive rates in sentencing. Therefore, imposing fairness constraints that equalize false

	Baseline		Mitigated (DP)		Mitigated (EOdd)		Mitigated (EOpp)	
	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced
Narcotics								
Δ_{DP}	0.097	0.035	0.027	0.009	0.028	0.027	0.053	0.016
Δ_{TP}	0.018	0.009	0.057	0.036	0.051	0.004	0.028	0.020
Δ_{FP}	0.145	0.013	0.082	0.010	0.078	0.012	0.104	0.007
Theft								
Δ_{DP}	0.141	0.064	0.024	0.039	0.014	0.029	0.054	0.062
Δ_{TP}	0.135	0.023	0.018	0.003	0.005	0.006	0.059	0.019
Δ_{FP}	0.136	0.134	0.021	0.068	0.012	0.050	0.039	0.127

Table 4. **Gradient Boosting:** Results for predictive analysis (**Charge reduction**) for baseline in comparison with Demographic parity (DP), Equalized odds (EOdd) and Equalized Opportunity (EOpp) mitigations. Here, again, Δ_{DP} denotes Demographic parity difference, Δ_{TP} the True positive difference and Δ_{FP} the False positive Difference.

	Baseline		Mitigated (DP)		Mitigated (EOdd)		Mitigated (EOpp)	
	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced
Narcotics								
Δ_{DP}	0.062	0.036	0.010	0.026	0.006	0.0001	0.012	0.027
Δ_{TP}	0.067	0.011	0.016	0.011	0.016	0.007	0.010	0.002
Δ_{FP}	0.065	0.083	0.012	0.065	0.004	0.020	0.025	0.072
Theft								
Δ_{DP}	0.103	0.027	0.004	0.009	0.024	0.004	0.007	0.025
Δ_{TP}	0.108	0.029	0.012	0.017	0.042	0.013	0.006	0.033
Δ_{FP}	0.079	0.077	0.022	0.030	0.010	0.001	0.023	0.077

Table 5. **Logistic Regression:** Results for predictive analysis (**Bond**) for baseline in comparison with Demographic parity (DP), Equalized odds (EOdd) and Equalized Opportunity (EOpp) mitigations. Here, again, Δ_{DP} denotes Demographic parity difference, Δ_{TP} the True positive difference and Δ_{FP} the False positive Difference.

positive rates across demographic groups is a promising avenue for mitigating sentencing disparities. Furthermore, our results suggest that the overreperesentation of black men in criminal case data contributes significantly to the observed biases. Mitigation approaches that recover the true distribution across demographic groups (e.g., in case precedents) could help to counteract such biases.

We again emphasize that this study does not aim on modeling the judicial decision process, but rather on identifying and understanding biases in legal data. Therefore, we do not evaluate the accuracy of our “typical” and “fair” judges with respect to the ground truth and focus solely on analyzing fairness metrics.

5 DISCUSSION

Guided by the question of *efficiently identifying and measuring biases in large-scale legal data*, we introduced a pipeline for the data-driven analysis of representation biases and sentencing disparities across demographic groups. We implemented two classical regression approaches and three commonly used fairness metrics to evaluate biases in criminal case data provided by the Cook County State's Attorney's Office. In particular, we presented four case studies, considering two

	Baseline		Mitigated (DP)		Mitigated (EOdd)		Mitigated (EOpp)	
	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced	Unbalanced	Balanced
Narcotics								
Δ_{DP}	0.076	0.074	0.004	0.013	0.005	0.029	0.052	0.085
Δ_{TP}	0.044	0.004	0.016	0.003	0.002	0.016	0.024	0.001
Δ_{FP}	0.124	0.164	0.019	0.001	0.007	0.020	0.096	0.181
Theft								
Δ_{DP}	0.026	0.051	0.023	0.022	0.020	0.031	0.026	0.048
Δ_{TP}	0.009	0.009	0.009	0.009	0.005	0.005	0.009	0.025
Δ_{FP}	0.053	0.104	0.047	0.046	0.046	0.051	0.053	0.115

Table 6. **Gradient Boosting:** Results for predictive analysis (**Bond**) for baseline in comparison with Demographic parity (DP), Equalized odds (EOdd) and Equalized Opportunity (EOpp) mitigations. Here, again, Δ_{DP} denotes Demographic parity difference, Δ_{TP} the True positive difference and Δ_{FP} the False positive Difference.

different crime types (narcotics and theft), as well as two different judicial decisions (charge reduction and free bonds). When comparing our baseline model, which represents the decisions of a “typical” judge according to the data, with that of a “fair” judge that applies one of three concepts of fairness, we observe a quantifiable difference in our fairness metrics.

The present study is, to our knowledge, the first to evaluate biases in legal data from an algorithmic fairness perspective. The results of our study echo general findings in previous quantitative studies [9, 21, 22], albeit focusing not on sentencing disparities in the final case outcome, but on other aspects of the legal process. Looking ahead, we hope that this study sheds some light on suitable notions of fairness for legal proceedings and on pipelines for systematically analyzing bias in large-scale legal data.

While the present study does not evaluate the accuracy of the judicial decisions of the “fair” and the “typical” judges, future work could look at the implications of fairness constraints on metrics such as the likelihood of repeated offenses (for charge reduction) or failure to appear in court (free bonds). Note that both approaches (the “fair” and the “typical” judge) aim to model average judicial decisions and do not account for individual differences between judges. For instance, the approach is not able to measure the impact of a few “unfair" judges in a larger group of mostly “fair" judges. We leave the study of such cases to future work.

Other directions for future work include a study of sentencing decisions, focusing on disparities in the final case outcome across demographic groups. This requires the adaption of the binary fairness metrics considered here to real-valued outputs. Furthermore, while the present study focused on fairness across demographic groups (*group fairness*), an interesting direction for future work would be the study of individual fairness notions. Finally, while much of the literature on biases in legal data has focused on criminal cases, we hope to study biases in civil cases too.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. 80 (10–15 Jul 2018), 60–69. http://proceedings.mlr.press/v80/agarwal18a.html

[2] Joseph J Avery and Joel Cooper. [n.d.]. Racial Bias in Post-Arrest and Pretrial Decision Making: The Problem and a Solution. *CORNELL JOURNAL OF LAW AND PUBLIC POLICY* 29 ([n. d.]), 257.

[3] Amanda Nicholson Bergold, Gregory Davis, Oana Dumitru, Asma Ghani, Rachel D Godsil, Rebecca C Hetey, Margaret Bull Kovera, Besiki Luka Kutateladze, Andrea Lyon, Naci Mocan, et al. 2020. Bias in the law : a definitive look at racial prejudice in the U.S. criminal justice system/ edited by Joseph Avery and Joel Cooper. (2020).

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (2021), 3–44. https://doi.org/10.1177/0049124118782533

[5] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[6] L. Elisa Celis, Vijay Keswani, and Nisheeth K. Vishnoi. 2020. Data preprocessing to mitigate bias: A maximum entropy based approach. arXiv:1906.02164 [cs.LG]

[7] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. https://doi.org/10.1089/big.2016.0047

[8] Maria-Veronica Ciocanel, Chad M. Topaz, Rebecca Santorella, Shilad Sen, Christian Michael Smith, and Adam Hufstetler. 2020. JUSTFAIR: Judicial System Transparency through Federal Archive Inferred Records. *PLOS ONE* 15, 10 (10 2020), 1–20. https://doi.org/10.1371/journal.pone.0241381

[9] Alma Cohen and Crystal S. Yang. 2019. Judicial Politics and Sentencing Decisions. *American Economic Journal: Economic Policy* 11, 1 (February 2019), 160–91. https://www.aeaweb.org/articles?id=10.1257/pol.20170329

[10] Briggs Depew, Ozkan Eren, and N. Mocan. 2017. Judges, Juveniles, and In-Group Bias. *The Journal of Law and Economics* 60 (2017), 209 – 239.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12).* Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[12] J.L. Eberhardt. 2019. *Biased: Uncovering the Hidden Prejudice That Shapes What We See, Think, and Do.* Penguin Publishing Group.

[13] K. Hao. 2019. AI is sending people to jail—and getting it wrong. *MIT Technology Review.* (2019). https://www.technologyreview.com/2019/01/21/137783/algorithms-criminal-justice-ai/

[14] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16).* 3323–3331.

[15] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R.* Springer. https://faculty.marshall.usc.edu/gareth-james/ISL/

[16] J. Kleinberg, H. Lakkaraju, J. Leskovec, Jens Ludwig, and S. Mullainathan. 2018. Human Decisions and Machine Predictions. *Economics of Networks eJournal* (2018).

[17] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2020. Algorithms as discrimination detectors. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30096–30100. https://doi.org/10.1073/pnas.1912790117

[18] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[19] David B. Mustard. 2001. Racial, Ethnic, and Gender Disparities in Sentencing: Evidence from the U.S. Federal Courts. *The Journal of Law and Economics* 44, 1 (2001), 285–314. https://doi.org/10.1086/320276

[20] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-Aware Data Mining *(KDD '08).* Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959

[21] M. Marit Rehavi and Sonja B. Starr. 2014. Racial Disparity in Federal Criminal Sentences. *Journal of Political Economy* 122, 6 (2014), 1320–1354. https://doi.org/10.1086/677255

[22] Max M. Schanzenbach and Emerson H. Tiller. 2007. Reviewing the Sentencing Guidelines: Judicial Politics, Empirical Evidence, and Reform. *University of Chicago Law Review* 75 (2007), 3.

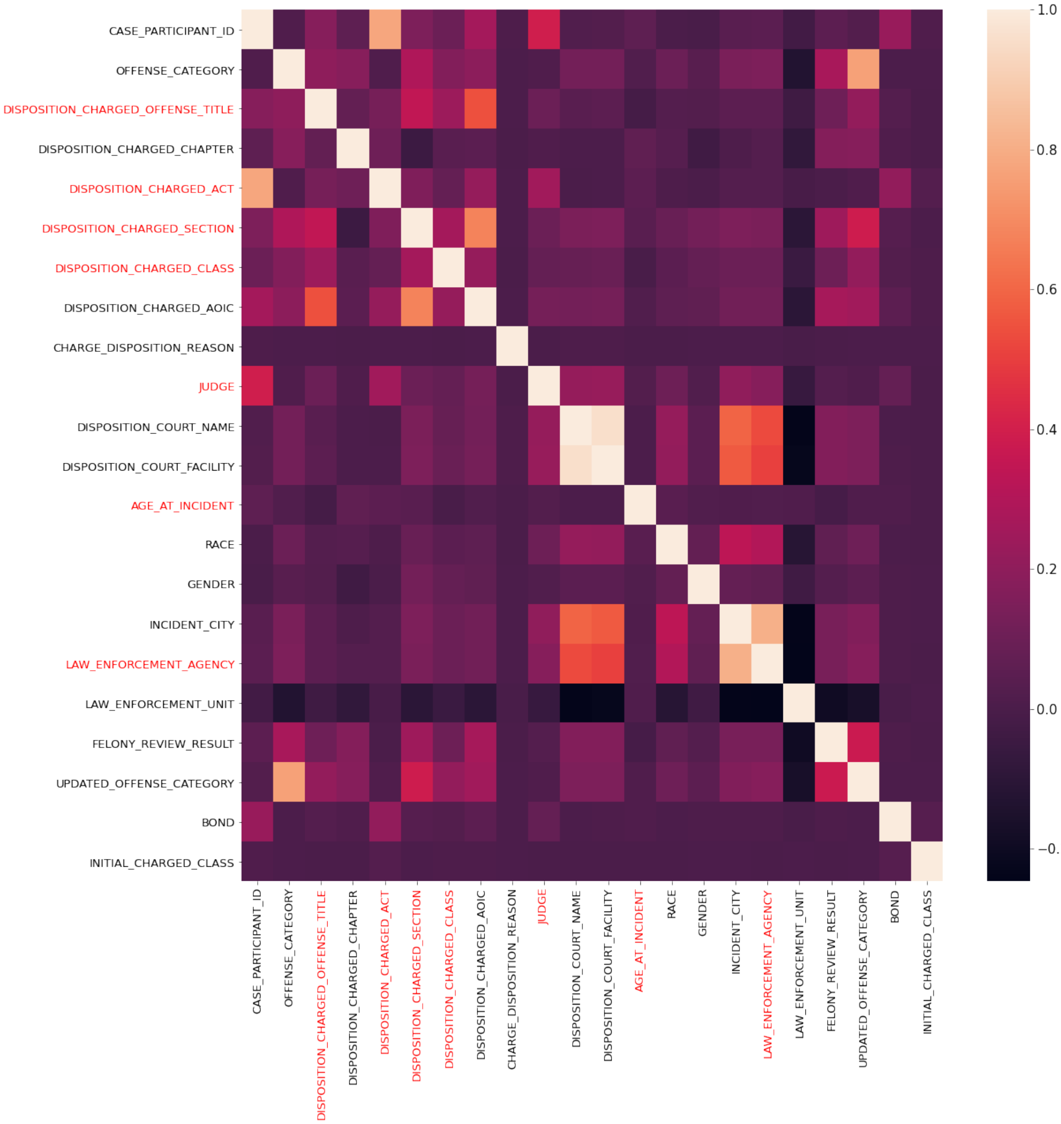

Fig. 5. Feature correlation in the post processed data. The features highlighted in red are used in our predictions. Lighter colors indicate a greater correlation.

A FEATURE SELECTION

Figure 5 shows the correlation between all features[6] in the preprocessed data. The following features were selected for the analysis:

[6] A glossary of all features with detailed descriptions can be found at https://datacatalog.cookcountyil.gov/api/views/7mck-ehwz/files/cec52aad-1f1a-4bf3-b02d-e5360298f66e?download=true&filename=CCSAO%20Data%20Glossary.pdf.

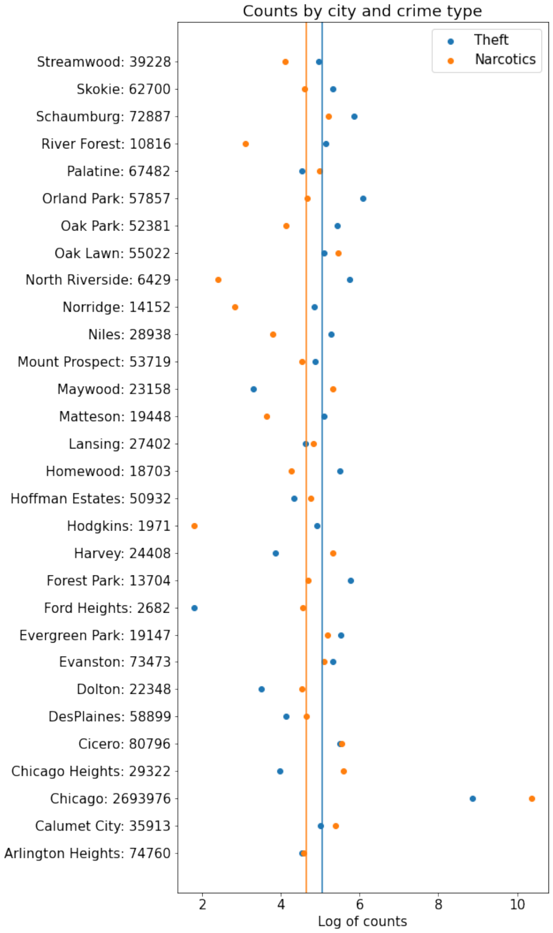

Fig. 6. Arrest counts across cities in Cook County by offense category. Cities are given with their estimated total population as of 2019, and arrest counts are given on a logarithmic scale. The lines represent the median count for each crime type.

- **Act, Section:** Legal act/ section for the charge according to the Illinois criminal statute. Both features provide information on the type of the crime and the context of the crime situation.
- **Age_At_Incident:** Age of defendant at the date of the incident as reported in the case intake.
- **Class:** Legal class for the charge.
- **Judge:** Judge who oversaw the case.
- **Law_Enforcement_Agency:** Law Enforcement agency associated with the arrest.
- **Disposition_Charged_Offense_Title:** Specific title of the charged offense at disposition.

B INCIDENT CITIES

Figure 6 shows arrest counts in 30 out of 142 incident cities in Cook County in the raw data. The shown data accounts for 89% of all arrests in Narcotics and Theft in Cook County included in the data set.

C MODELING INFORMATION

C.1 Hyperparameters

Model	Parameters

Logistic Regression	max_iter = 5000
	class_weight = 'balanced'

Gradient Boosting	Default

Table 7. The hyperparameters used to train our two types of models. All Gradient Boosting or unspecified Logistic Regression parameters are left as sklearn's default.

C.2 Class Frequencies and Weights

	Frequency		Weight	
	True	False	True	False
Narcotics				
Charge Reduction	0.484	0.516	1.033	0.969
Bond	0.576	0.424	0.867	1.180
Theft				
Charge Reduction	0.508	0.492	0.984	1.017
Bond	0.683	0.317	0.732	1.580

Table 8. Class frequencies and weights across sentencing features and crime types. The weights are those used in our weighted logistic regression model, and calculation specifics can be found at https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html