Blind Justice: Algorithmically Masking Race in Charging Decisions

Alex Chohlas-Wood
alexcw@stanford.edu
Stanford University

Joe Nudell
jnu@stanford.edu
Stanford University

Zhiyuan "Jerry" Lin
zylin@cs.stanford.edu
Stanford University

Julian Nyarko
jnyarko@law.stanford.edu
Stanford University

Sharad Goel
scgoel@stanford.edu
Stanford University

ABSTRACT

A prosecutor's decision to charge or dismiss a criminal case is a particularly high-stakes choice. There is concern, however, that these judgements may suffer from explicit or implicit racial bias, as with many other such actions in the criminal justice system. To reduce potential bias in charging decisions, we designed a system that algorithmically redacts race-related information from free-text case narratives. In a first-of-its-kind initiative, we deployed this system at a large American district attorney's office to help prosecutors make race-blind charging decisions, where it is used to review many incoming felony cases. We report on both the design, efficacy, and impact of our tool for aiding equitable decision making. We demonstrate that our redaction algorithm is able to successfully mask race-related information, making it difficult for a human reviewer to guess the race of a suspect. In the jurisdiction we study, we found little evidence of disparate treatment in charging decisions even prior to deployment of our intervention. Thus, as expected, our tool did not substantially alter charging rates. Nevertheless, our study demonstrates the feasibility of race-blind charging, and, more generally highlights the promise of algorithms to bolster equitable decision making in the criminal justice system.

1 INTRODUCTION

It is a staple of a fair judicial system that justice should be blind. Indeed, the image of Justitia, who's "most enigmatic" trait is the blindfold covering her eyes as she passes judgment [28], is almost ubiquitous [16]. For centuries and across cultures, the notion of a blind umpire has been held up as a depiction of a fair and equitable decision maker. In the context of criminal justice, the ideal of "blind justice" transforms into a specific, normative prescription: In deciding on a defendant's fate, members of the judicial branch and law enforcement should not take into account immutable features. For a smaller subset of these features, the normative prescription evolves into a constitutional command. A decision motivated by a consideration of so-called "protected characteristics," such as a defendant's race or gender, constitutes a violation of the Equal Protection Clause of the Fourteenth Amendment [30].[1] Yet ample evidence suggests that immutable features—even those enjoying constitutional protection—regularly influence the criminal process at various stages, whether it be in the context of policing, prosecution, detention, adjudication or sentencing [18, 35, 59]. Consistent with these findings, a majority of U.S. adults believes that the criminal justice system systematically favors white over Black suspects.[2]

A particularly critical decision point in the criminal process is the prosecutorial decision whether or not to charge a case. Prosecutors enjoy much discretion in deciding who and on what grounds to prosecute [14], a discretion that Justice Jackson once called "the most dangerous power of the prosecutor" [27]. The legal bar to substantiate a claim that this discretion has been misused ("selective prosecution") is notoriously high [41]. Indeed, the first and last time the Supreme Court concluded that the enforcement of a criminal law violated the Equal Protection Clause was in 1886 with *Yick Wo v. Hopkins*. In its most recent case from 1996, *Armstrong*, the Court effectively made it impossible for Black defendants to compel evidence for selective prosecution from the government unless the defendants can already demonstrate that "similarly situated" suspects were not prosecuted. At the same time, observational studies suggest that discretion in prosecutorial charging decisions may significantly contribute to racial disparities [58].

The picture painted by the empirical evidence, paired with the high hurdles presented to those seeking legal recourse, often raises the question whether justice "is really blind" [18, 52] or whether the idea of blind justice is merely a "myth" [56]. Actually blinding decision makers, while previously contemplated [49], has so far evaded serious scholarly consideration, likely because it was thought to be infeasible.[3]

In this paper, we report on a pilot initiative that does just that. We implemented an algorithm at the district attorney's office of a major U.S. county to automatically remove race-related information from incident reports of felony cases before they are reviewed by a prosecutor. In many jurisdictions—including the one we study—a prosecutor's decision to charge a case is largely based on a written police report, conversations with officers, and, in some cases, review of photographic and other physical evidence. To facilitate race-blind charging, we used methods from natural language processing to mask several types of information in the incident reports that could indicate an individual's race, including: explicit mentions

[1] Through reverse incorporation, the same standards apply to the federal government, including federal prosecutors, under the Fifth Amendment. *See* Boiling v. Sharpe, 347 U.S. 497, 499 (1954).

[2] https://www.pewsocialtrends.org/2019/04/09/race-in-america-2019/

[3] Legal scholars have primarily discussed normative aspects of blind decision making in the context of the "John Doe" defendant, i.e. when the defendant cannot be identified or should not be identified because the charges levied against him, even if unjustified, carry with them a social stigma [64].

Stanford University, October 5, 2020, Stanford, CA
2020.

of race; select physical descriptors, including hair and eye color; individuals' names or nicknames; location information, including neighborhood names and street addresses; and officer names, given that prosecutors may remember where officers are stationed.

In order to verify our tool, we assessed the ability of an expert annotator to infer the race of the individuals described in the masked narratives. When asked to estimate the likelihood that a Black individual was involved in a case, we found that the annotator achieved an AUC of 0.70 [95% CI: 0.64–0.76], on par with the performance of a supervised model (0.75 [95% CI: 0.71–0.78]) trained on the masked text. We further found that these results were roughly comparable to the AUC (0.63 [95% CI: 0.58–0.68]) of a model based only on the alleged offenses. Because it is important for prosecutors, at a minimum, to see the allegations, this similarity suggests our algorithm is able to mask narratives close to the practical limit.

To aid use of our masking algorithm, we implemented a new two-stage procedure for case review. First, before speaking with officers or reviewing any non-redacted information on a case, a prosecutor makes (and records) a preliminary charging decision based solely on the algorithmically redacted incident report. Next, the prosecutor reviews all the available evidence on the case, including an unredacted incident report, and makes a final charging determination. If, however, the preliminary and final decisions differ, prosecutors are required to provide a written explanation for the change. This new procedure required transitioning the office from a primarily paper-based system to an internal web-based platform that we created. We designed this process and platform to limit the role of race in charging decisions while also preserving the opportunity for a full review of all relevant case information.

We used a quasi-experimental approach to investigate the impact of our algorithm on charging decisions. We found that masking did not substantially alter overall charging rates. We further found that race-specific charging rates were similar for masked and unmasked case files. However, even prior to adoption of our masking algorithm, we found little evidence of racially disparate treatment in our partner jurisdiction, based on an observational analysis of historical charging decisions. Our results thus provide further evidence that race does not substantially impact charging decisions in the jurisdiction we examine.

We caution, however, that these findings should not be interpreted as proof for the absence of discrimination in charging decisions more broadly. For one, our study is limited in geography. Observational studies suggest that biases in prosecutorial charging decisions may be a more significant problem in other districts [58]. We hypothesize our intervention could yield greater effects elsewhere. In addition, with its focus on selective prosecution, our work is tailored to detecting and mitigating disparities caused by differences in the perception of race, which is merely one form of discrimination. Our analysis does not investigate whether prosecutorial charging decisions lead to differential burdens across racial lines, which is another form of discrimination recognized under the law.[4] Finally, it bears emphasis that our results focus not on biases in the entire criminal process, but only on a single decision point.

[4]For instance, Title VII of the Civil Rights Act of 1964 regonizes "disparate impact" as a form of discrimination, which, among others, applies to facially neutral hiring policies that disproportionately burden minorities.

Our work demonstrates that blind decision making can be facilitated through the use of computational methods. The feasibility of our implementation prompts the need for a serious and concrete debate surrounding the normative aspects of blinding decision makers at the different stages of the criminal process. In particular, the more widespread use of technology such as ours in the context of criminal justice can have ambivalent effects on public trust. On one hand, it may increase the confidence in the criminal justice system and, for that reason, may constitute a beneficial intervention even in districts such as ours that show no evidence of selective prosecution. On the other, many people are averse towards algorithmic or algorithm-assisted decision making [15] and the knowledge that computational tools are involved in the decision-making process could increase that aversion. It remains unclear how these public concerns should be balanced off against the advantages of blinding.

2 BACKGROUND AND RELATED WORK

Our study contributes to a growing literature on algorithms in the law. Algorithms have become an increasingly central aspect of the legal environment. In the realm of private law, they are used in many commercial contexts [61], such as contract drafting [6, 9] and contract review [26, 36, 37]. They are also increasingly prevalent in the analysis of consumer contracts [24], and in discovery during civil litigation [66]. On the administrative side, a recent report finds that nearly half of 142 examined federal agencies have implemented machine learning tools [17], reaching from predictive enforcement, to facial and handwriting recognition, to automatic adjudicatory error correction.

In criminal law, the vast majority—if not all—of commonly used algorithms are designed to predict some future outcome that has the potential to inflict significant costs or benefits on the individual [12, 13]. For instance, the prediction of a defendant's recidivism risk can be the difference between pretrial detention and release. These types of algorithms raise two important legal and normative challenges. The first set of concerns pertains to the forward-looking nature of predictive algorithms. Because these algorithms tend to penalize individuals not for conduct that they have done, but for conduct that they may do in the future, scholars have argued that the use of predictive assessments is inconsistent with our existing theories of punishment, most importantly with principles of retributive justice [23, 44]. In addition, there is concern about the potential for false positives that is inherent in predictions [40, 42].

The second set of challenges focuses on the source of information that most predictive algorithms use. Generally speaking, the performance of predictive algorithms improves as they receive more information as an input. Accurate prediction algorithms often use information about group characteristics (e.g., gender or socioeconomic groups) to predict the behavior of an individual. This effectively conditions punitive elements on the mere fact that the defendant is a member of a particular group, a generalization that raises important concerns under anti-discrimination laws [57]. Does the inclusion of protected characteristics (e.g., race or gender) automatically constitute a form of unconstitutional discrimination? Does this prohibition extend to socioeconomic characteristics? How should one treat attributes that are not themselves susceptible, but strongly correlate with protected class characteristics (e.g., zip code)? As the use of predictive algorithms becomes more ubiquitous, their interaction with anti-discrimination laws remains at the center of the legal debate [21, 31, 32, 50].

Our design avoids many of these normative challenges by breaking with the tradition of how algorithms are used in the criminal justice system. In contrast to the forward-looking algorithms currently employed that focus on predicting an individual's future behavior, our algorithm to mask racial information can be conceived of as a mechanism that enables the implementation of an otherwise difficult redaction process. In principle, it is possible to employ humans to remove racial identifiers from incident reports, a practice that is unlikely to raise serious concerns about appropriate theories of punishment. However, as this practice would be prohibitively costly and performance may fluctuate from one human to the next, our algorithmic approach offers a reliable and economically feasible alternative.

Similarly, our implementation does not raise the same normative concerns under anti-discrimination laws. Rather than relying on group-based information (including protected characteristics, like race) as inputs, we aim to remove information from the text of incident reports in order to effectively mask racial cues. Indeed, whereas the goal of previous implementations to maximize predictive performance is often in tension with anti-discrimination doctrine, the goal of our system is to help decision makers avoid engaging in conduct that is prohibited by anti-discrimination laws.

In addition to its contributions to the literature on algorithms in law, our study makes a unique contribution to the extensive literature seeking to analyze the influence of race on different decision points in the criminal process [2]. Many studies focus on disparities in sentencing for Black and white defendants [18, 43, 47, 60, 62]. Other work has considered the role of race in policing [7, 19, 46], arrests [4] and plea-bargaining [5]. More recently, scholars have attempted to assess the cumulative effect of race across multiple decision points [33, 35]. While the majority of these studies suggests that people of color are disadvantaged, the evidence is often ambiguous and sensitive to the specific methodology applied and criminal context under investigation [18, 62].

Although it did not escape scholarly attention, fewer studies have focused on the importance of race in the prosecutorial charging decision [11, 34, 51, 63]. Perhaps the most ambitious examination to date was conducted by Starr and Rehavi [58]. Examining 36,659 individuals in the federal criminal justice system from the initial arrest to final sentencing, the authors found that the primary driver for sentencing disparities between Black and white defendants stem from differences in the initial charging decision of the prosecutor, specifically for charges with statutory mandatory minimum sentences. In contrast, a recent experimental study by Robertson et al. [48] found no evidence of racial biases in charging decisions. The authors presented prosecutors with vignettes in which the race of the suspect was randomly varied and asked, among others, whether the prosecutors would press charges. In an observational analysis of prosecutors at the San Francisco District Attorney's Office, MacDonald and Raphael [39] similarly found little evidence of disparate treatment in charging decisions. A few studies even found that prosecutors may exert biases in favor of minority suspects [65].

Overall, the evidence on race effects in charging decisions, like the evidence at other decision points, is ambiguous. Part of this ambiguity may be explained by differences in geography or crime type of the data under study [34]. In addition, some inquiries suffer from small sample sizes [11]. We hypothesize that yet another reason for the observed variability in results is driven by methodological differences across the studies. Indeed, in order to identify racial effects, nearly all previous efforts have relied on the "selection on observables" assumption. That is, the researcher tries to adjust for characteristics that are correlated with both race and charging decisions (such as alleged crime type), and assumes that the residual variation can be causally attributed to the race of the suspect. However, as is well known, this approach fails if there are unobserved characteristics that may confound the results. For instance, a study that adjusts for broad categories of crime may still be unable to take into account variation in the severity of the alleged crime within any particular category [58]. In the few instances in which researchers have tested race effects in an experimental setting [48], the hypothetical nature of the experiment can make it difficult to extrapolate the findings to the real world.

Our study is the first to directly manipulate the perception of race for charging decisions in the district attorney's office of a major U.S. county. Although practical considerations prevented us from conducting a perfect randomized controlled trial (see Section 4 for a discussion of these limitations), our design circumvents the most serious concerns with purely observational approaches, complementing the results of past work and suggesting a path forward for future researchers.

3 METHODOLOGY

To reduce the role of race in charging decisions, we made three substantial changes to the case review and charging process at our partner district attorney's office. First, we developed a redaction algorithm which automatically identifies and redacts race-related words. Next, we designed a new two-stage case review procedure that incorporates blind review while preserving attorney discretion with all available case information. Finally, we built a custom web-based (private intranet) platform to display these redacted case narratives and record the case decisions made by prosecutors.

3.1 Masking narratives

Our blinding algorithm automatically identifies race or race-related information in the free-text narrative included in every incident report. We identify and obscure five types of information: (1) explicit mentions of race; (2) select physical descriptors, including hair and eye color; (3) individuals' names or nicknames; (4) location information, including neighborhood names and street addresses; and (5) officer names, given that prosecutors may remember where officers are stationed.

Many of these types of information are identified through the use of predefined regular expressions. For example, we match against an openly available dataset of every street and neighborhood name in our partner jurisdiction to identify instances of location information. However, these identifications must be made with care to avoid over-redaction (e.g., to avoid matching "Main Street" to every mention of "main"). Similarly, we must avoid redacting color descriptions when

Original narrative

Lucy Johnson reported that a Black male with brown hair wearing a black jacket assaulted her in Midtown, next to Johnson's home. She reported the incident to Officer Lee.

Automatically redacted narrative

[Victim 1] reported that a [race] male with [hair color] wearing a black jacket assaulted her in [neighborhood], next to [Victim 1]'s home. She reported the incident to [Officer 1].

Figure 1: Automated redactions from a fictional narrative excerpt. Mentions of race, physical descriptors, names, and locations are all identified and re-labeled to preserve readability. Additionally, person labels are enumerated to track each person's role across a case. Non-race-related descriptions (like "black jacket") are preserved.

they do not refer to racial labels (e.g., "Black male" vs. "black car"). Our filters adapt to these circumstances by specifying additional criteria for a match. For example, to identify mentions of streets, we require that the matching street name be preceded by a number, or followed by a street type, such as "road" or "boulevard".

We identify individual names using a combination of both regular expressions and named entity recognition—a technique to automatically locate and classify mentions of people, places, and other "named entities" in unstructured text. Each incident report includes a structured list of individual participant names, which we leverage to identify instances of names in the narrative. Named entity recognition assists this process by identifying other possible mentions of names which are not exact matches to the list of involved individuals. Named entity recognition is also particularly helpful when individual names are entirely omitted in the list of involved individuals.

Complete obfuscation of race and race-related terms—as might be visually implied by the black-bar redaction common in the release of federal documents—would make the narrative largely incomprehensible to a prosecutor. For example, with black-bar redaction, an attorney may be unable to distinguish between the actions of the victim and suspect in an assault case. Our algorithm preserves this information by indicating the type of information obscured, and enumerates mentions of each individual, so that they have the same label wherever they are referenced. Figure 1 provides an example. We note that certain demographic information is not redacted because it can have direct bearing on whether a prosecutor decides to charge a case (e.g., a victim's gender may inform whether a physical altercation was mutually provoked or more likely one-sided).

3.2 Blind review procedure and platform.

Coupled with the masking algorithm, we designed a two-stage case review process to limit the role of race in charging decisions while also preserving the opportunity for a full review of all relevant case information. In particular, our procedure addresses a concern that redacted or omitted information may significantly impair a prosecutor's charging decision. For example, given that it is difficult to imagine useful and effective redaction of certain evidence—including photo, video, or audio—we do not include these sources when asking an attorney to make a race-blind case decision. But these sources may reasonably influence a prosecutor's charging decision. To address this concern, we require prosecutors to review any given case twice: first, they conduct a preliminary blinded review with our redacted incident report; and later, they engage in an expanded, final review with all available (unredacted) information.

The initial race-blind review occurs the first workday morning after an individual has been booked into jail, before a prosecutor has spoken with officers or reviewed any non-redacted information on a case. At this stage, prosecutors review only a limited set of case information: date and time information; basic information about all involved individuals, such as sex, age, height, and weight; details on confiscated property; categorical flags, such as whether the incident was gang-related; a list of proposed charges; and our redacted case narrative. After reviewing this information, the prosecutor is asked to select one of four options for the likely final charging decision: "charge", "probably charge", "probably discharge", and "discharge". Prosecutors are also asked to explain their decision with a brief comment.[5]

At the charging decision deadline, typically 1–2 days later, the same prosecutor conducts a final comprehensive review of the case. This review includes not only the full, unredacted version of the incident report, but also photo or video evidence, and any supplementary reports filed in the interim. The prosecutor reviews all these documents and makes a charging decision on the case. Crucially, if this decision differs from their initial, race-blind determination, they must explain the reason for the change. This stage of the process is intended to encourage prosecutors to pause and consider why they have chosen to change their decision compared to the initial, race-blind review.

This new procedure required transitioning the office from a primarily paper-based system to an internal web-based platform that we created. Both algorithmic redaction and preliminary review take place on our new platform. We plan to open source this software after testing its extension to several other district attorney's offices.

4 RESULTS

4.1 Assessing redaction quality

To statistically evaluate the quality of our redactions, we took two approaches. To begin, we asked an independent researcher—with previous experience reading unredacted narratives from this jurisdiction—to read the algorithmically redacted narratives and then predict whether a Black suspect was involved.[6] In preparation for the labeling task, the annotator saw the race of involved individuals on 15 redacted reports. Then we recorded their predictions in two separate ways. First, we asked the annotator to assess each redacted narrative individually, and to provide a probability estimate for whether a Black suspect was involved in the incident.

[5] In the event that redaction was insufficient or erroneous, prosecutors can also leave feedback on the quality of the redaction process, or explain other possible reasons they could not review the case.

[6] There were separate reasons we could not complete the same task for Hispanic or Asian suspects. First, Hispanic suspects are rarely classified as such in our partner jurisdiction. For Hispanic results in this paper, we resort to using Census name frequencies [46] to estimate the likelihood that a given individual is Hispanic, but this process makes the label partly unreliable. Second, there are too few Asian suspects in our sample to reliably measure any difference in performance.

Because providing precise probability estimates can be difficult for humans, we complemented the approach with a second task. We asked the annotator to review a pair of redacted narratives, informing them that exactly one of the narratives involved a Black suspect; the task was to identify the incident involving the Black suspect.

As an additional method of evaluation, we trained a series of machine-learning models to infer race using varying levels of case information. One model had access to the same information as our human reviewer—namely, the redacted narrative and the basic case information provided to the prosecutor during blind review. The redacted narrative was represented by transforming individual words into word embeddings using a 300-dimension GloVe embedding. To obtain document embeddings, we then averaged over the word embeddings. We also explicitly counted the occurrence of each redaction token (e.g., "[race/ethnicity]") and included these counts as features. A second model was trained to infer race using only the alleged charge, which we assume to be the minimal necessary information required to make a sensible charging decision. We would expect that perfectly effective race redaction which preserves charge information would perform no better than this simple model, making it a suitable baseline. A third model had access to every case detail available (except race itself), including the unredacted narrative. We excluded race from this model to measure the need for algorithmic redaction beyond the simple removal of a suspect's race from the incident report. For these prediction tasks, we trained several gradient-boosted decision tree models—a state-of-the-art machine-learning method for such classification problems. Model performance was assessed using ten-fold cross validation. A fourth model simply guesses the most common class every time to gauge the performance of a naive model which does not use any information.

This task was assessed on the approximately 400 cases reviewed during the pilot by our partner jurisdiction. In this sample, approximately 43% of incidents involve a Black suspect, 30% involve a Hispanic suspect, 29% involve a white suspect, 4% involve an Asian suspect, and 3% involve a suspect of another race.[7] For the sake of brevity, we report here on each model's ability to predict whether a Black individual was listed as a suspect, though we note the full findings (listed in Table 1) are qualitatively similar when predicting the presence of individuals from other racial or ethnic groups.

The baseline model—using only charge information—achieves an AUC of 0.63 [95% CI: 0.58–0.68] when predicting the presence of a Black suspect. The full model achieves an AUC of 0.92 [95% CI: 0.90–0.93]. The human annotator with access to the redacted narratives achieves an AUC of 0.65 [95% CI: 0.58–0.72] and an AUC of 0.70 [95% CI: 0.64–0.76] for the single-case and pair-wise predictions, respectively. The annotator slightly outperformed the baseline model, but the difference is not statistically significant.

Overall, our results suggest that the unmasked incident reports include many identifiers that allow for the identification of a suspect's race. However, redaction successfully conceals racial cues, making it difficult for our human annotator to acquire additional information about the suspect's race. We do note that, with an AUC of 0.75 [95% CI: 0.71–0.78], our classifier outperforms our human annotator when given the redacted incident report. We thus cannot rule out that a prosecutor would be able to pick up on cues that our human annotator misses. However, it was neither practically feasible nor normatively desirable to amplify a prosecutor's ability to identify racial cues through our annotation exercise.

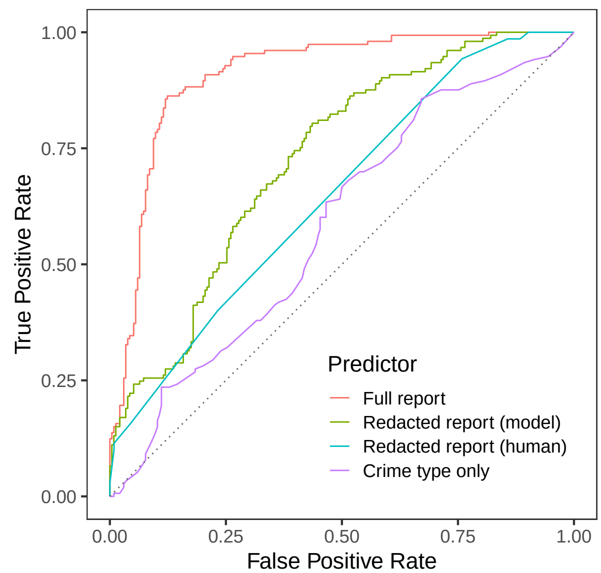


Figure 2: ROC curves for models with access to varying levels of case information. Models are evaluated on their ability to predict whether a Black suspect was involved in each incident. Better performance is indicated by curves that approach the upper left corner; the dotted diagonal line indicates a baseline model that guesses race completely at random. Both the human and model with access to redacted information performs comparably to the baseline crime type model, while the model with access to all information performs substantially better than both. This suggests our blind charging algorithm is able to effectively redact race-related information from incident reports.

4.2 Impact on charging decisions

Next we evaluate the impact of our deployment on charging practices. We aimed to randomize redaction at the level of individual cases, but a conventional randomized controlled trial was not feasible due to operational considerations. We randomly selected up to 20 felony cases every day for blind review. However, as a practical necessity, the unit supervisor had discretion to reassign cases. Further, attorneys were encouraged but not required to carry out a blind review of the cases they were assigned. As a result, cases that ultimately underwent blind review were comparable—but not statistically equivalent—to those that did not undergo blind review. Prosecutors have conducted over 400 preliminary blind reviews since pilot initiation in August 2019. Figure 3 plots the number of

[7] A single incident can involve multiple suspects and so these proportions sum to more than 100%.

Table 1: Auditing redaction efficacy. Lower and upper estimates correspond to a 95% confidence interval. Note that the Hispanic label is partly imputed from individual's names by classifying an individual as Hispanic if Census records for that first and last name show it is commonly associated with individuals of Hispanic ethnicity. As a result, Hispanic predictions using the full model can replicate the label perfectly (and are thus invalid), given the presence of the same attribute in the feature set.

	Naive	Crime type		Redacted		Full	
	Acc.	Acc.	AUC	Acc.	AUC	Acc.	AUC
Black	61% [61%–61%]	60% [58%–63%]	0.63 [0.58–0.68]	67% [63%–72%]	0.75 [0.71–0.78]	87% [85%–90%]	0.92 [0.9–0.93]
Hispanic	72% [72%–73%]	73% [68%–78%]	0.7 [0.64–0.75]	76% [74%–79%]	0.71 [0.68–0.75]	– –	– –
White	73% [73%–74%]	73% [69%–78%]	0.55 [0.49–0.62]	73% [70%–77%]	0.69 [0.62–0.77]	83% [80%–87%]	0.89 [0.87–0.91]
Crime type		✓		✓		✓	
Age, sex, date				✓		✓	
Categorical flags				✓		✓	
Redacted embeddings				✓			
Unredacted embeddings						✓	
Location						✓	
Census race inference from names						✓	

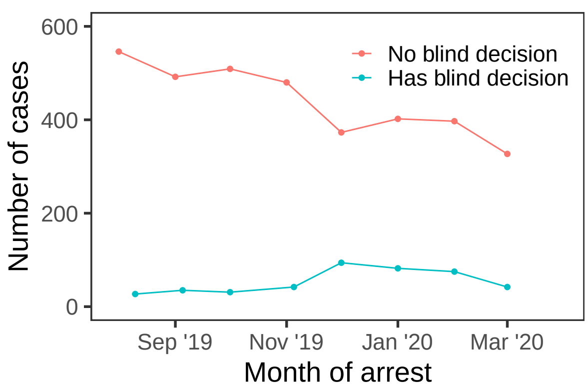


Figure 3: Cases over time. The blue line represents cases with a completed blind review; every other felony case without a completed blind review is depicted in red.

cases with and without a blind review during the period of examination. As can be seen, the majority of cases are decided without blinding, reflecting the limitations just discussed.

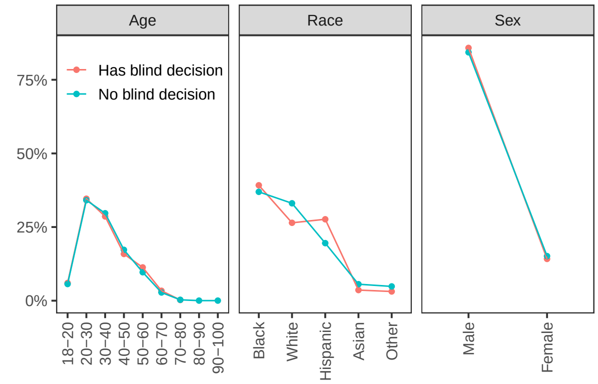
In Figures 4a and 4b, we assess the balance of several of our covariates between cases that underwent blind review. The balance plots suggests cases were largely randomly assigned for blind review, alleviate concerns of selection bias due to human discretion. One exception is the narcotics case-type, which represents a crime category that has a small number of highly specialized prosecutors who receive most of these cases. We nonetheless opt to take a conservative approach and view and analyze our intervention as a quasi-experimental design, as described below.

Table 2: Actual charging rates, by race. Margins of error represent 95% confidence intervals. Note that—in generating these raw statistics—we did not adjust for factors that partially explain observed differences.

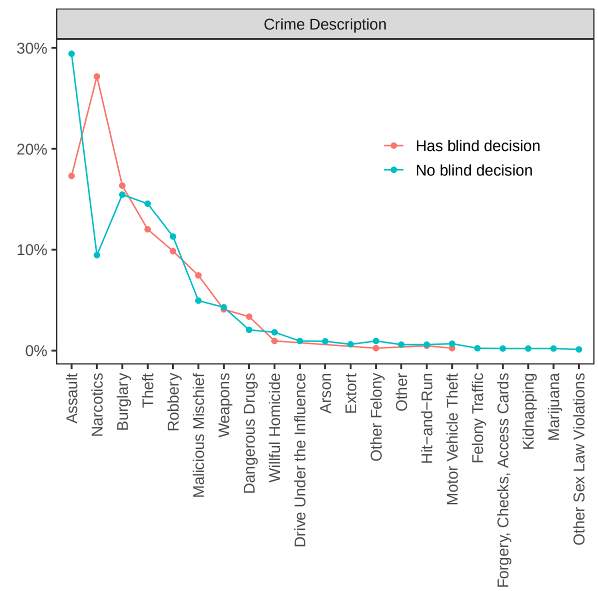
	Blinding	**No blinding**	**Overall**
Asian	47% [21–72%]	53% [46–60%]	52% [46–59%]
Black	52% [44–60%]	56% [53–59%]	56% [53–58%]
Hispanic	68% [59–76%]	51% [47–55%]	54% [50–57%]
White	55% [45–64%]	49% [46–52%]	50% [47–52%]
Overall	57% [52–61%]	52% [50–54%]	53% [51–54%]

In addition to limitations in our randomization strategy, we note that case-level treatment assignment might suffer from spillover effects. For example, the use of redaction on some cases could impact decisions on non-redacted cases by drawing attention to various elements of the non-redacted case files. In theory, one could avoid potential spillover effects by randomizing assignment at the level of prosecutors rather than cases. However, in our partner jurisdiction, a small number of prosecutors make all initial charging decisions, and so randomization at the level of the prosecutor would significantly diminish statistical power.

To assess the pilot's impact on charging practices, we compare cases with a completed blind review to those without. Roughly 57% (95% CI: [52–61%]) of cases with a blind review were eventually charged, compared to 52% (95% CI: [50–54%]) of cases without a blind review. Table 2 breaks these numbers down by race. For example, 52% (95% CI: [44–60%]) of redacted cases involving Black individuals were charged, compared to 56% (95% CI: [53–59%]) of unredacted cases. The small sample sizes—particularly for redacted



(a) Demographic variables.



(b) Crime description.

Figure 4: Balance plots comparing select attributes between cases that did and did not receive a blind review.

cases—make it difficult to precisely estimate charging rates. However, the observed differences, both overall and across racial groups, are generally small.

As discussed above, this finding is consistent with observational studies that have found little evidence of disparate treatment in the jurisdiction we consider. To reproduce these findings, we modeled charging decisions for arrests that occurred in the six years prior to the start of the pilot. We estimated charging rates as a function of race, sex, and age; the day, month, and year of the arrest; the presence of flags on the incident report indicating domestic violence, elderly victims, gang involvement, weapons, or the use of a body-worn camera; the Census-derived racial composition of the area in which the incident occurred; the precinct where the arrest occurred; two-year retrospective arrest and felony arrest counts for the suspect; the alleged charges; and the number of alleged charges

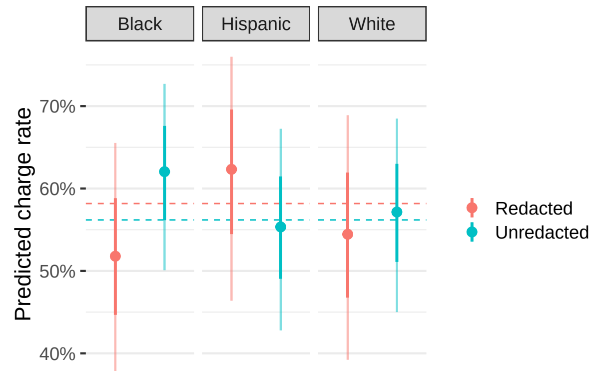


Figure 5: Estimated charging rates for a canonical individual: a 35-year-old male arrested on a Monday in February on a single assault charge. We note that—after adjusting for all listed covariates—charging rates are roughly equivalent across race groups.

in total. After adjusting for these factors, we found no statistically significant difference in charging rates for Black and Hispanic suspects relative to white suspects. These results confirm that disparate treatment in the charging decision is likely not a significant factor in creating racial disparities in our partner jurisdiction.

As noted earlier, case assignment was not perfectly randomized and it is likely that some of the factors that caused decisions to be made on the platform correlate with charging decisions. For instance, a prosecutor who is more likely to use the platform may also specialize in a particular type of crime. To address the problem of possible confounding, we supplement the examination of the raw charging rates in Table 2 with an analysis that seeks to adjust for observed differences in cases. In particular, we fit a logistic regression to estimate the relationship between blind review, race, and charging decision while adjusting for several possible confounders, including: the suspect's age, sex, and local arrest and charge histories; date and location of arrest; a fixed effect for each reviewing prosecutor; and the alleged crime type(s) (Table 3). We also interact race with our indicator for whether a case was redacted to allow for the effect of redaction to vary across racial groups.

After adjusting for these factors, we again find no statistically significant difference in charging rates between cases with a blind review and those without, although the relatively small number of cases make it difficult to estimate the effect precisely. Specifically, we estimate that cases which received a blind review had 0.9 times the odds of being charged as those without a blind review, with a 95% CI of 0.6–1.4. To illustrate how effects vary by race and blinding, Figure 5 depicts estimated charging rates for a hypothetical suspect—a 35-year-old man arrested on a Monday in February with a single assault charge—under varying assumptions about his perceived race. We see generally similar—though imprecisely estimated—charging rates across race groups, both when race-related information is redacted and when it is not.

Finally, we examine the preliminary decisions of prosecutors, based solely on their read of the redacted incident report. In their

Table 3: Selected logistic regression coefficients, presented on the odds scale, for a model of charging decisions made during the pilot period. Lower and upper estimates correspond to a 95% confidence interval. White individuals who did not receive a blind review are treated as the baseline.

	Simple	Demographic	Full	Full w/ Pros.
Redacted	1.2 [0.8–1.8]	1.2 [0.8–1.8]	1.1 [0.7–1.7]	0.9 [0.6–1.4]
Asian	1.2 [0.9–1.6]	1.2 [0.9–1.6]	1.3 [0.9–1.8]	1.5 [1.0–2.1]
Asian x Redacted	0.6 [0.2–1.9]	0.6 [0.2–1.7]	0.5 [0.2–1.7]	0.5 [0.1–1.6]
Black	1.3 [1.1–1.6]	1.3 [1.1–1.5]	1.3 [1.1–1.5]	1.2 [1.0–1.5]
Black x Redacted	0.7 [0.4–1.1]	0.7 [0.4–1.2]	0.7 [0.4–1.2]	0.7 [0.4–1.3]
Hispanic	1.1 [0.9–1.3]	1.0 [0.8–1.2]	1.0 [0.8–1.2]	0.9 [0.7–1.2]
Hispanic x Redacted	1.6 [0.9–2.9]	1.6 [0.9–2.9]	1.4 [0.8–2.6]	1.5 [0.8–2.8]
Other	0.8 [0.6–1.1]	0.8 [0.5–1.1]	0.8 [0.5–1.1]	0.8 [0.5–1.1]
Other x Redacted	0.9 [0.3–2.9]	0.9 [0.3–2.9]	0.5 [0.2–1.9]	0.6 [0.2–2.1]
Race	✓	✓	✓	✓
Redacted	✓	✓	✓	✓
Age, Sex		✓	✓	✓
Day/Month of Arrest			✓	✓
Precinct			✓	✓
2-yr ct. of arrests			✓	✓
2-yr ct. of filed felonies			✓	✓
Crime type			✓	✓
Prosecutor fixed effect				✓

Table 4: Number of cases, split by preliminary and final decision. For simplicity, we have grouped the possible preliminary decisions "probably charge" and "charge" into a single category (with a similar consolidation for dismissed cases).

	Actually charged	**Actually dismissed**
Blind charge	187	95
Blind dismiss	49	85

initial decisions, as seen in Table 4, prosecutors recommended pressing charges in 68% (95% CI: 63%–72%) of cases. After reviewing the complete, unredacted case file, charging rates drop to 57% (95% CI: 52%–61%). We note two possible mechanisms for the lower final charging rate. First, there may be a de-personalization effect from the platform, where the lack of personal information obscured by redaction causes prosecutors to act more punitively. Alternatively, it is possible that prosecutors overestimate the likelihood that the full, unredacted case files would contain incriminating evidence.

5 DISCUSSION

In the context of criminal justice, algorithms are most commonly employed to impose punitive measures in the ostensible service of public safety (e.g., as with predictive policing and pretrial risk assessment). That history may explains popular discontent with algorithmically aided decisions. But, as our implementation shows, algorithms can also be used to reign in potential abuses of power. As modern data analytics have helped to police the police [22], so too can algorithms help assess and rectify the actions of other decision makers in the criminal justice system.

More broadly, we are seeing a new type of algorithm emerge; one that is designed at the outset to protect the rights and support the needs of system-involved individuals. For example, recent work uses reinforcement learning to craft personalized text message reminders for individuals with upcoming court dates, and optimization methods to offer select individuals transportation vouchers to further improve court appearance rates [1]. This latest generation of supportive algorithms—which aims to reduce social stratification—creates new challenges for anti-discrimination doctrine. The current focus of the legal community lies on constraining the input factors for algorithmic decision making. While particularly important for predictive algorithms with punitive consequences, this focus provides little guidance to those who seek to use algorithms in the service of furthering due process goals. For example, if an algorithmic system learns that men and women, or Black and white individuals, respond differently to court appointment reminders, should that information be used to design better personalized communications? Likewise, should protected characteristics be used to statistically inform the allocation of limited transportation benefits? It remains unclear how to trade off concerns arising out of the use of protected characteristics against the desire for tailored interventions that seek to support the recipient.

In our particular application—race blind charging—it is similarly unclear what characteristics ought to be masked from the decision maker. If we mask race, should we also redact information about other protected classes, such as gender and religious affiliation? Beyond legally protected classes, should masking extend to socioeconomic information? What if the decision maker seeks to favor a specific minority group and blinding prevents the protected group from enjoying a more favorable treatment?[8] Further, while blinding can help counteract selective prosecutorial practices, it can also remove information that is important to determine the exact circumstances and likely culpability of the alleged conduct. For instance, racial cues may inform prosecutors about the likelihood that an infringement on the victim's rights was racially motivated and may thus constitute a hate crime. As the applications of algorithms continue to evolve, it will be of crucial importance to assess and evaluate these and related questions on the interaction of technology and anti-discrimination laws.

Blinding is one form of remedy aimed at removing disparities between members of different racial categories, but there are alternatives. For instance, rather than not using racial information at all, Butler [8] supports the use of affirmative action in the criminal process, which promulgates the use of race-conscious policies to the benefit of racial minorities (e.g., by mandating that Black suspects cannot be subjected to criminal enforcement of drug-related crimes in a way that is disproportionate to their actual rate of commission). This raises the question whether blind decision making, albeit feasible, is also normatively most desirable. In particular, blinding necessarily removes identifying and individualizing information from the decision making process, potentially making it more difficult to both empathize or feel anger towards the individuals involved [54, 55]. Although there is evidence to suggest that deidentification can alter punitiveness [55], scholars have been resistant to discuss the role empathy ought to play in the context of adjudication and law enforcement [10]. The imminence of tools that allow for widespread deidentification bring new urgency to these debates and make their associated costs and benefits more concrete.

In addition to the potential for our blinding algorithm to improve prosecutorial behavior, it offers a new path for auditing decisions. In practice, as we have discussed above, it can be difficult to redact case details in a true randomized controlled trial. Still, we believe the general methodology we propose is a compelling and complementary alternative to traditional observational analysis and experiments in synthetic environments.

In the jurisdiction we study, our proposed methodology reveals no clear evidence for racial biases in prosecutorial charging decisions. We highlight, however, that the scope of our study is limited in several important aspects. First, with approximately 4,000 cases in total, and with under 500 blind reviews, our sample size is relatively small, making it difficult to obtain precise estimates. Second, our results do not necessarily extend to other jurisdictions, as past observational studies suggest that selective prosecution may be a more significant problem in other districts.

Third, the racial biases we study merely cover one aspect of impermissible conduct recognized under anti-discrimination laws. But even in the absence of such biases, prosecutorial charging decisions can impose undue burdens on racial minorities that can be equally susceptible to concerns about discrimination [3, 20, 29]. Consider, for instance, criminal processing in the context of drug enforcement. Prosecutors are less likely to use their discretion to the benefit of the suspect if the incident involves crack cocaine as opposed to powdered cocaine [25]. At the same time, the share of minorities among those reporting to have used crack cocaine at some point in their life is significantly higher than for powdered cocaine [45]. Hence, a practice that is more likely to press charges for use of crack cocaine than powdered cocaine impacts racial minorities disproportionately [53], even in the absence of selective prosecution across racial lines. It is important to emphasize that our findings do not speak to the presence or significance of such disparities in the effect of prosecutorial discretion.

Finally, we focus only on prosecutorial charging decisions, an early step in the criminal process. While it is important to obtain causal estimates for racial biases at every decision point [20], a fuller picture of the role of race in criminal justice requires one to consider a multitude of steps, such as racially motivated policing, dismissal and sentencing [35].

Our work highlights the rapidly expanding development and use of algorithms in criminal justice, both for auditing and for improving behavior. Whereas past work has largely focused on statistical risk prediction, there is urgent need for a more comprehensive legal and normative debate surrounding the broader class of algorithms now emerging. In addressing the accompanying concerns, legal scholars can provide a robust foundation that will help orient, refine and appropriately constrain the development of new tools as they are introduced into the criminal justice system.

[8] Previous research suggests that prosecutors may favor female suspects with reduced charging decisions [38], and so blinding towards gender could increase charges for women.

REFERENCES

[1] Amy Adams. 2020. Stanford Impact Labs forges partnerships to tackle social problems. *Stanford News* (2020). https://news.stanford.edu/2020/07/07/stanford-impact-labs-forges-partnerships-tackle-social-problems/

[2] Joseph Avery and Joel Cooper. 2020. *Bias in the Law: A Definitive Look at Racial Prejudice in the U.S. Criminal Justice System.* Rowman & Littlefield. Google-Books-ID: z6nSDwAAQBAJ.

[3] Ian Ayres. 2005. Three tests for measuring unjustified disparate impacts in organ transplantation: the problem of "included variable" bias. *Perspectives in Biology and Medicine* 48, 1 Suppl (2005), S68–87.

[4] Shima Baradaran. 2013. Race, Prediction, and Discretion. *George Washington Law Review* 81, 1 (2013), 157–222.

[5] Carlos Berdejó. 2018. Criminalizing Race: Racial Disparities in Plea-Bargaining. *Boston College Law Review* 59 (2018), 1187–1249.

[6] Kathryn D. Betts and Kyle R. Jaep. 2017. The Dawn of Fully Automated Contract Drafting: Machine Learning Breathes New Life Into a Decades-Old Promise. *Duke Law & Technology Review* 15, 1 (2017), 216–233.

[7] Anthona Y. Braga, Rod K. Brunson, and Kevin M. Drakulich. 2019. Race, Place, and Effective Policing. *Annual Review of Sociology* 45 (2019), 535–555. https://doi.org/10.1146/annurev-soc-073018-022541

[8] Paul Butler. 1997. Affirmative Action and the Criminal Law. *University of Colorado Law Review* 68, 4 (1997), 841–890.

[9] Anthony J. Casey and Anthony Niblett. 2017. Self-Driving Contracts. *The Journal of Corporation Law* 43 (2017), 1–33.

[10] Thomas B. Colby. 2012. In Defense of Judicial Empathy. *Minnesota Law Review* 96 (2012), 1944–2015.

[11] Katy M. Colon, Philip R. Kavanaugh, Don Hummer, and Eileen M. Ahlin. 2018. The impact of race and extra-legal factors in charging defendants with serious sexual assault: Findings from a five-year study of one Pennsylvania court jurisdiction. *Journal of Ethnicity in Criminal Justice* 16, 2 (2018), 99–116. https://doi.org/10.1080/15377938.2018.1439791

[12] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[13] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), 797–806. http://doi.org/10.1145/3097983.3098095

[14] Angela J. Davis. 1998. Prosecution and Race: The Power and Privilege of Discretion. *Fordham Law Review* 67, 1 (1998), 13–68. https://heinonline.org/HOL/P?h=hein.journals/flr67&i=29

[15] Berkeley J Dietvorst, Joseph P. Simmons, and Cade Massey. 2015. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal*

of Experimental Psychology: General 144, 1 (2015), 114–126.
[16] Doron Dorfman. 2016. The Blind Justice Paradox: Judges with Visual Impairments and the Disability Metaphor. *Cambridge International Law Journal* 5, 2 (2016), 272–305. https://www.elgaronline.com/view/journals/cilj/5-2/cilj.2016.02.05.xml
[17] David F. Engstrom, Daniel E. Ho, Catherine M. Sharkey, and Mariano-Florentino Cuéllar. 2020. *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies: Report Submitted to the Administrative Conference of the United States*. Technical Report. https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf
[18] Travis W. Franklin. 2018. The State of Race and Punishment in America: Is Justice Really Blind? *Journal of Criminal Justice* 59 (2018), 18 – 28. https://doi.org/10.1016/j.jcrimjus.2017.05.011
[19] Roland G. Fryer J. 2018. An Empirical Analysis of Racial Differences in Police Use of Force. *Journal of Political Economy* (2018), forthcoming.
[20] Johann Gaebler, William Cai, Guillaume Basse, Ravi Shroff, Sharad Goel, and Jennifer Hill. 2020. Deconstructing Claims of Post-Treatment Bias in Observational Studies of Discrimination. (2020). Working paper.
[21] Talia B. Gillis and Jann L. Spiess. 2019. Big Data and Discrimination. *The University of Chicago Law Review* 86 (2019), 459–487.
[22] Sharad Goel, Maya Perelman, Ravi Shroff, and David Alan Sklansky. 2017. Combatting police discrimination in the age of big data. *New Criminal Law Review: An International and Interdisciplinary Journal* 20, 2 (2017), 181–232.
[23] Bernard E. Harcourt. 2006. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* University of Chicago Press, Chicago, Illinois.
[24] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous
[25] Richard D. Hartley, Sean Maddan, and Cassia C. Spohn. 2007. Prosecutorial Discretion: An Examination of Substantial Assistance Departures in Federal Crack-Cocaine and Powder-Cocaine Cases. *Justice Quarterly* 24, 3 (Sept. 2007), 382–407. https://doi.org/10.1080/07418820701485379
[26] Fahad ul Hassan and Tuyen. 2020. Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. *Journal of Legal Affairs and Dispute Resolution in Engineering and Construction* 12, 2 (2020). https://doi.org/10.1061/(ASCE)LA.1943-4170.0000379
[27] Robert H. Jackson. 1940. The Federal Prosecutor. *Journal of Criminal Law and Criminology* 31, 1 (1940), 3–6.
[28] Martin Jay, Costas Douzinas, and Lynda Nead. 1999. Must Justice Be Blind? In *Law and the Image.* The University of Chicago Press, Chicago and London, 19–35.
[29] Jongbin Jung, Sam Corbett-Davies, Ravi Shroff, and Sharad Goel. 2018. Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651* (2018).
[30] Pamela S. Karlan. 1998. Race, Rights, and Remedies in Criminal Adjudication. *Michigan Law Review* 96 (1998). https://repository.law.umich.edu/mlr/vol96/iss7/2
[31] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Cass R. Sunstein. 2018. Discrimination in the Age of Algorithms. *Journal of Legal Analysis* 10 (Dec. 2018), 113–174. https://doi.org/10.1093/jla/laz001
[32] Joshua Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, David G. Robinson, and Harlan Yu. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165 (2017), 633–705.
[33] Megan C. Kurlychek and Brian D. Johnson. 2019. Cumulative Disadvantage in the American Criminal Justice System. *Annual Review of Criminology* 2 (2019), 291–319. https://doi.org/10.1146/annurev-criminol-011518-024815
[34] Besiki Kutateladze, Vanessa Lynn, and Edward Liang. 2012. *Do Race and Ethnicity Matter in Prosecution? A Review of Empirical Studies.* Technical Report. Vera Institute of Justice. https://www.vera.org/publications/do-race-and-ethnicity-matter-in-prosecution-a-review-of-empirical-studies
[35] Besiki L. Kutateladze, Nancy R. Andiloro, Brian D. Johnson, and Cassia C. Spohn. 2014. Cumulative Disadvantage: Examining Racial and Ethnic Disparity in Prosecution and Sentencing. *Criminology* 52, 3 (2014), 514–551. https://doi.org/10.1111/1745-9125.12047
[36] LawGeex. 2018. Comparing the Performance of Artificial Intelligence to Human Lawyers in the Review of Standard Business Contracts. (2018). https://images.law.com/contrib/content/uploads/documents/397/5408/lawgeex.pdf
[37] Jeehee Lee, June-Seong Yi, and JeongWook Son. 2019. Development of Automatic-Extraction Model of Poisonous Clauses in International Construction Contracts Using Rule-Based NLP. *Journal of Computing in Civil Engineering* 33, 3 (2019). https://ascelibrary.org/doi/10.1061/%28ASCE%29CP.1943-5487.0000807
[38] Zhiyuan Jerry Lin, Alex Chohlas-Wood, and Sharad Goel. 2019. Guiding Prosecutorial Decisions with an Interpretable Statistical Model. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society.* 469–476.
[39] John MacDonald and Steven Raphael. 2017. *An Analysis of Racial and Ethnic Disparities in Case Dispositions and Sentencing Outcomes for Criminal Cases Presented to and Processed by the Office of the San Francisco District Attorney.* Technical Report. University of Pennsylvania.
[40] Michael Marcus. 2009. MPC—the Root of the Problem: Just Deserts and Risk Assessment. *Florida Law Review* 61 (2009), 751–776. http://smartsentencing.com/Marcus_FLRev9-09.pdf
[41] Richard H. McAdams. 1998. Race and Selective Prosecution: Discovering the Pitfalls of Armstrong. *Chicago-Kent Law Review* 73 (1998), 605–667.
[42] Pari McGarraugh. 2013. Up or Out: Why "Sufficiently Reliable" Statistical Risk Assessment Is Appropriate at Sentencing and Inappropriate at Parole. *Minnesota Law Review* 97 (2013), 1079–1113. https://scholarship.law.umn.edu/cgi/viewcontent.cgi?article=1351&context=mlr
[43] Ojmarrh Mitchell. 2005. A Meta-Analysis of Race and Sentencing Research: Explaining the Inconsistencies. *Journal of Quantitative Criminology* 21, 4 (2005), 439–466. http://doi.org/10.1007/s10940-005-7362-7
[44] John Monahan. 2006. A Jurisprudence of Risk Assessment: Forecasting Harm among Prisoners, Predators, and Patients. *Virginia Law Review* 92, 3 (2006), 391–435. https://www.virginialawreview.org/sites/virginialawreview.org/files/391.pdf
[45] Joseph J. Palamar, Shelby Davies, Danielle C. Ompad, Charles M. Cleland, and Michael Weitzman. 2015. Powder cocaine and crack use in the United States: An examination of risk for arrest and socioeconomic disparities in use. *Drug and Alcohol Dependence* 149 (April 2015), 108–116. https://doi.org/10.1016/j.drugalcdep.2015.01.029
[46] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, and Sharad Goel. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour* (2020), 1–10. https://doi.org/10.1038/s41562-020-0858-1
[47] Travis C. Pratt. 1998. Race and Sentencing: A Meta-Analysis of Conflicting Empirical Research Results. *Journal of Criminal Justice* 26, 6 (1998), 513–523.
[48] Christopher Robertson, Shima Baradaran Baughman, and Megan S. Wright. 2019. Race and Class: A Randomized Experiment with Prosecutors. *Journal of Empirical Legal Studies* 16, 4 (2019), 807–847.
[49] Sunita Sah, Christopher T. Robertson, and Shima B. Baughman. 2015. *Blinding Prosecutors to Defendants' Race: A Policy Proposal to Reduce Unconscious Bias in the Criminal Justice System.* Technical Report. Behavior Science & Policy Association.
[50] Matthew U. Scherer, Allan King, and Marko Mrkonich. 2020. Applying Old Rules to New Tools: Employment Discrimination Law in the Age of Algorithms. *South Carolina Law Review* 71 (2020), forthcoming. https://ssrn.com/abstract=3472805
[51] Lauren O'Neill Shermer and Brian D. Johnson. 2010. Criminal Prosecutions: Examining Prosecutorial Discretion and Charge Reductions in U.S. Federal District Courts. *Justice Quarterly* 27, 3 (2010), 394–430. http://doi.org/10.1080/07418820902856972
[52] Jonathan Simon. 2019. Explicit Bias: Why Criminal Justice Reform Requires Us to Challenge Crime Control Strategies That Are Anything But Race Blind. *Tulsa Law Review* 54, 2 (2019), 331–338.
[53] David A. Sklansky. 1994. Cocaine, Race, and Equal Protection Essay. *Stanford Law Review* 47, 7 (1994), 1283–1322. https://heinonline.org/HOL/P?h=hein.journals/stflr47&i=1309
[54] Deborah A. Small and George Loewenstein. 2003. Helping a Victim or Helping the Victim: Altruism and Identifiability. *Journal of Risk and Uncertainty* 26, 1 (Jan. 2003), 5–16. https://doi.org/10.1023/A:1022299422219
[55] Deborah A. Small and George Loewenstein. 2005. The devil you know: the effects of identifiability on punishment. *Journal of Behavioral Decision Making* 18, 5 (2005), 311–318. https://doi.org/10.1002/bdm.507
[56] Harold J. Spaeth, David B. Meltz, Gregory J. Rathjen, and Michael V. Haselswerdt. 1972. Is Justice Blind: An Emprical Investigation of a Normative Ideal. *Law & Society Review* 7, 1 (1972), 119–137. http://www.jstor.org/stable/3052832
[57] Sonja B. Starr. 2014. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66, 4 (2014), 803–872. http://www.stanfordlawreview.org/wp-content/uploads/sites/3/2014/04/66_Stan_L_Rev_803-Starr.pdf
[58] Sonja B Starr and M. Marit Rehavi. 2014. Racial Disparity in Federal Criminal Sentences. *Journal of Political Economy* 122 (2014), 1320–1354. https://repository.law.umich.edu/articles/1414
[59] Lisa Stolzenberg, Stewart J. D'Alessio, and David Eitle. 2013. Race and Cumulative Discrimination in the Prosecution of Criminal Defendants. *Race and Justice* 3, 4 (2013), 275–299. https://doi.org/10.1177/2153368713500317
[60] Laura T. Sweeney and Craig Haney. 1992. The Influence of Race on Sentencing: A Meta-Analytic Review of Experimental Studies. *Behavioral Sciences and the Law* 10 (1992), 179–195.
[61] George Triantis. 2013. Improving Contract Quality: Modularity, Technology, and Innovation in Contract Design. *Stanford Journal of Law, Business & Finance* 18 (2013), 177–214.
[62] Jeffrey Ulmer. 2018. *Race, Ethnicity, and Sentencing.* Oxford University Press. https://oxfordre.com/criminology/view/10.1093/acrefore/9780190264079.001.0001/acrefore-9780190264079-e-262
[63] Jeffrey T. Ulmer, Megan C. Kurlycheck, and John H. Kramer. 2007. Prosecutorial Discretion and the Imposition of Mandatory Minimum Sentences. *Journal of*

Research in Crime and Delinquency 44, 4 (2007), 427–458. http://doi.org/10.1177/0022427807305853
[64] Michael S. Vogel. 2004. Unmasking John Doe Defendants: The Case against Excessive Hand-Wringing over Legal Standares. *Oregon Law Review* 83, 3 (2004), 795–860. https://heinonline.org/HOL/P?h=hein.journals/orglr83&i=805
[65] John Wooldredge and Amy Thistlethwaite. 2004. Bilevel Disparities in Court Dispositions for Intimate Assault. *Criminology* 42, 2 (2004), 417–456. https://doi.org/10.1111/j.1745-9125.2004.tb00525.x
[66] Eugene Yang, David Grossman, Ophir Frieder, and Roman Yurchak. 2017. Effectiveness Results for Popular e-Discovery Algorithms. *ICAIL '17: Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law* (2017), 261–264. http://dx.doi.org/10.1145/3086512.3086540