From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

Doctoral Dissertations, 2011-2026 Psychology

2021

Legal Data: Bias in the Law, and How Legal Technology Can Be Built to Help Correct For It

Avery, Joseph

Princeton, NJ : Princeton University

http://arks.princeton.edu/ark:/88435/dsp013j333534t

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

LEGAL DATA:

BIAS IN THE LAW, AND

HOW LEGAL TECHNOLOGY CAN BE BUILT TO HELP CORRECT FOR IT

Joseph J. Avery

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PSYCHOLOGY

Adviser: Joel Cooper

May 2021

© Copyright by Joseph J. Avery, 2021. All rights reserved.

Abstract

There is significant racial disparity in U.S. incarceration rates, with African Americans vastly overrepresented in prisons and jails. This was true a century and a half ago, and it is true today. Surprisingly, and in spite of decades of social psychological research, we still do not have clear answers to fundamental questions undergirding this fact: what accounts for the racially disproportionate rates of incarceration, and at what points in the criminal justice process does bias emerge, causing minority defendants to be treated differently than White defendants?

In this dissertation, I cover three aspects of this problem. First, I consider bias in the law, focusing on the criminal justice actors most central to plea bargaining: criminal defense attorneys and prosecutors. I also focus on their decision making during a critical period: post-arrest and pretrial. Second, I consider stereotypes of criminal subtypes and what these stereotypes mean for legal decision making. Third, I propose that the most feasible and promising approach is to guide criminal justice actors' decision making through the use of machine-generated outputs. Across theoretical and empirical work, I outline the case for this approach and a potential psychological difficulty that may stand in the way of making it a reality.

Table of Contents
with Summaries of Dissertation Chapters

P. 1	**Part I: Is there Bias in the Law? Analyses of Key Criminal Justice Actors**	
P. 3	Chapter 1	**Implicit and Explicit Attitudes in Criminal Defense Attorneys** Criminal defense attorneys practicing in forty-three U.S. states completed both implicit and explicit measures designed to capture racial bias. On average, the criminal defense attorneys surveyed held White preference implicit bias. In addition, when the defendants were Black, the more implicit bias the attorneys held, the worse was their interpersonal regard for the defendants.
P. 28	Chapter 2	**Implicit Bias, Fear of Being Labeled Biased, and Criminal Defense Attorneys' Motivation to Correct for Bias in the Law** Criminal defense attorneys were surveyed regarding implicit and explicit bias in themselves and their peers. They also were asked about their fears of being labeled biased, and they were given the opportunity to formulate a strategy for overcoming bias in the law. Our results suggest that, if one wants to know how much effort a defense attorney will expend on racial equality in the law, one should ask three questions. First, is the attorney implicitly biased? Second, does he or she worry about being accused of bias? Third, does the attorney believe that racial bias impacts criminal defense decision making? There was an observed correlation between, one, the attorney's bias attorney; two, how much the attorney worried about being accused of bias; and, three, the outcome measure of effort expended on legal racial equality, such that increases in the first two correlated with a drop in effort. On a promising note, the more the attorney believed in the effect bias has on criminal defense decision making, the more effort he or she expended.
P. 43	Chapter 3	**Bias in Prosecutorial Decision Making** I statistically evaluated charge reductions in a large, urban prosecutor's office, determining which factors were driving charge reductions, including whether race was an important factor in the same. The findings consistently showed disparate rates of charge reductions, with White defendants (vs. Black and Hispanic defendants) benefiting. Even after controlling for a nearly exhaustive list of factors, race still was driving these case outcomes.
P. 57	**Part II: What Drives Bias in Legal Outcomes?**	
P. 59	Chapter 4	**Identifying Stereotypes of Criminal Subtypes via Reverse Correlation**

		Criminal subtype stereotypes have long been a vexing subject. There is a lack of consensus regarding both the content and the source of the stereotypes. In this study, we used reverse correlation to produce visual images of individuals' stereotypes of criminal subtypes, and we used these results to proffer a new understanding of what drives the stereotypes. While criminal subtype stereotypes are certainly influenced by crime base rates and individual experiences, among many factors, they are significantly driven by stereotypes of immorality. We found a Black/White parsing in criminal subtype stereotypes that occurred along the moral dimension. That is, *malum in se* offenses, generally deemed immoral, were associated with Black individuals, while *malum prohibitum* offenses, generally deemed wrong but not necessarily immoral, were associated with White individuals.
P. 79	Chapter 5	**How Imputations of Immorality Color Criminal Subtype Stereotypes and Impact Case Outcomes** I shed light on how the morality-driven stereotypes discussed in Chapter 4 help explain an inconsistency that has plagued research regarding the effects of stereotypes on case outcomes; specifically, research regarding crime congruency. The race-crime congruency effect is often found with stereotypically Black crimes but not with stereotypically White crimes. I provide evidence that Black defendants are stereotyped as immoral actors, and this has a dramatic effect on case outcomes. For guilt determinations, this imputation of immorality overwhelms the crime congruency effect, thus leading to the observed Black/White imbalance. In short, the theory helps to explain racially disparate rates of both guilt adjudication and sentence impositions.
P. 99	**Part III: Can Legal Technology Be Used to Decrease Bias in the Law?**	
P. 101	Chapter 6	**Attitudes Regarding Bias in Legal Technology** I explore perceptions of legal technology and, more specifically, how bias in legal technology is perceived and assessed. I ask, is there reverse automation bias for legal technology? If there is, does it threaten the adoption of technology that shows promise for decreasing racial bias in legal outcomes? I hypothesized that, given biased outcomes, machine adjudicators will be trusted less than human adjudicators. In addition, more drastic corrective action will be recommended in response to machine bias (vs. human bias). The five studies employ similar prompts, all of which describe either a human legal decision maker (a prosecutor) or a machine decision maker, that is, an algorithm trained to automate prosecutors' decision making in plea bargaining. I explore participants' trust ratings for these adjudicators (the algorithm and the prosecutor)

		when racial disparities are discovered, and I also explore participants' response in terms of trying to correct for the bias. While occurrences of bias in humans lead to decreased trust in the humans, occurrences of bias in machines completely eroded trust in the machines. Moreover, I found that this attitudinal backlash has potentially grave consequences: machine decision makers that were as adept at legal decision making tasks and twice as racially fair as human decision makers were still trusted less than the human decision makers.
P. 133	Chapter 7	**Decreasing Attorney Discretion through Algorithmic Decision Making—Examples from Criminal and Civil Law** I outline a method and process for how attorneys can use their own historical data to build computational models that enable them to anchor on racially fair machine generated outcomes. The driving insight is that, while we typically think of artificial intelligence in terms of automation (replacing human decision making), we instead should think of it as a second opinion, one that raises red flags and corrects human decisions that are straying towards bias. Using actual criminal case data, I provide a simulation of how this method would work in criminal law. Then, I discuss the development of an artificial legal intelligence that predicts the outcomes of civil legal cases, and I outline how it might be employed to decrease bias.
P. 174	**Concluding Thoughts**	
P. 175	**References**	
P. 218	**Appendix**	

Part I

Is there Bias in the Law? Analyses of Key Criminal Justice Actors

More than 2 million people are incarcerated in the United States (Sawyer & Wagner, 2020). A disproportionate number of individuals incarcerated in the U.S. are Black (Burke & Leben, 2007; Wagner, 2012). In 2015, the U.S. population was 13% Black (U.S. Department of Commerce, 2015) while the U.S. state prison population was 38% Black (Carson, 2015). Wagner (2012) found that Black individuals are more than four times as likely to be incarcerated as White individuals. One in 3 Black males will go to prison during their lifetimes, while the same is true for only 1 in 17 White males (Bonczar, 2003; see also Carson, 2014). In 1871, Frederick Douglass, speaking in Washington, D.C. to an audience that included United States President Ulysses S. Grant, said, "It is a real calamity, in this country, for any man, guilty or not guilty, to be accused of crime, but it is an incomparably greater calamity for any colored man to be so accused" (Douglass, 1871/1969, p. 133-134). It is clear that the calamity that Douglass described nearly 150 years ago remains a calamity to the present day. In a 2015 speech, then-President Obama cited the disproportionate number of minority men in prison in concluding that the criminal justice system was broken (Baker, 2015). Today, there is widespread, bipartisan agreement with Obama's conclusion (Young & Petersilia, 2016).

When it comes to such disparities, the post-arrest and pretrial (PAPT) period looms above the rest of the criminal justice process in importance. It is here that almost all criminal cases resolve (Pastore & Maguire, 2003; Durose & Langan, 2007; Motivans, 2015). In the federal system, only 2% of criminal defendants go to trial, and in the past two decades, the number of federal defendants opting for a trial has fallen by 60% (Gramlich, 2019). In addition, two actors—criminal defense attorneys and prosecutors—play central roles in the PAPT period, each

on an opposing side of the adversarial process. Given this, studying the attitudes of these actors is crucial in order to understanding the causes of racially disparate outcomes. In the first part of this dissertation, I focus on these actors. Do they hold racial bias? If so, what are the downstream consequences: does such bias matter for legal outcomes?

Chapter 1

Implicit and Explicit Attitudes in Criminal Defense Attorneys

Bias in the law is a well-studied area, but it nonetheless presents as something of a puzzle: there is no single factor that explains all racial disparities in criminal justice, and the extent to which different factors have contributed is not well understood. While attention has been paid to potential bias in police, prosecutors, jurors, and judges (Avery & Cooper, 2020a), one important group has gone relatively unstudied: criminal defense attorneys (Richardson & Goff, 2013). Ossei-Owusu (2019) has referred to this as a "scholarly gap." For example, *The Washington Post* recently compiled a list of all the studies they could find on bias in the criminal justice system (Balko, 2020). They divided the list into categories, including ones for police, jurors, prosecutors, judges, prison management, and even school discipline—and there was no category for criminal defense.

In this chapter, I examine these overlooked actors and this neglected aspect of the criminal justice process. I suggest that members of the defense team—criminal defense attorneys—harbor implicit racial bias toward Black defendants. Moreover, I suggest that those biases have consequences, as they may impact a key element of criminal defense practice, namely, attorneys' interpersonal regard for their clients.

Bias in Criminal Defense

Empirical research on criminal justice disparities has generally focused on the extent to which legislators (Alexander, 2010), law enforcement (Correll et al., 2007; Eberhardt, Goff, Purdie, & Davies, 2004), prosecutors (Kutateladze, Andiloro, & Johnson, 2016), jurors (Levinson & Young, 2010), judges (Rachlinski, Johnson, Wistrich, & Guthrie, 2009), and parole officers (Huebner & Bynum, 2008) hold implicit or explicit prejudice toward Black individuals.

By contrast, very little attention has been directed to the role played by defense attorneys—the people whose advice, given that almost all U.S. cases resolve through plea bargaining,[1] is critical in determining defendants' outcomes. Whether defense attorneys contribute to unequal outcomes is an important but understudied area.

Probably the most extensive empirical research on implicit bias in criminal defense was conducted by Eisenberg and Johnson (2004), who surveyed 238 capital defense lawyers.[2] The lawyers took a pencil-and-paper race Implicit Association Test (IAT), and the findings showed White preference implicit bias by White attorneys. Because the authors used a paper and pencil version of the IAT, however, they were limited in the extent to which they could make claims about the strength or magnitude of the bias in absolute terms. Our work seeks to extend the initial findings of Eisenberg and Johnson by assessing the correlates of such implicit racial bias and how the bias might affect the quality of representation.

Edkins (2011) uncovered racially disparate outcomes in defense attorneys' practice. She presented 95 criminal defense attorneys, about half of whom were practicing in the state of Florida, with vignettes in which she had manipulated race, and she found that plea bargains recommended for the African American client contained more severe sentences than did those recommended for the White client. Though this study presents important evidence suggesting that defense attorneys may offer worse representation to African American clients, there are two caveats worth noting: first, the sample size arguably was small and, second, the study relied on a race manipulation alone and did not include a racial attitude measure.

[1] According to the Bureau of Justice Statistics, for every 100 felony defendants processed in urban courts, only three are convicted at trial (Cohen & Kyckelhahn, 2010). In the federal system, only 2% of criminal defendants go to trial, and in the past two decades, the number of federal defendants opting for trial has fallen by 60% (Gramlich, 2019).
[2] In particular, the sample included 92 Georgia trial lawyers who were involved in representing defendants charged with capital crimes and 146 lawyers who represented death row inmates in federal habeas corpus proceedings.

There is a need for further research into criminal defense attorneys' implicit bias as a factor in the criminal justice process. Scholars have speculated that implicit bias drives unequal treatment of criminal defendants by their own attorneys (see Blakemore, 2016; Lyon, 2012). In their *Yale Law Journal* article, Richardson and Goff (2013) put the point plainly, calling for partnerships between social psychologists and criminal defense attorneys so that the claim of bias might be proved or disproved. They wrote, "[W]ithout data collection, it is simply impossible to know whether similarly situated clients are being treated alike" (Richardson & Goff, 2013, pp. 2648-2649).

The Defense Attorney-Client Relationship

Unlike prosecutors, jurors, and judges, criminal defense attorneys occupy a unique position in relation to defendants: they work together and develop interpersonal relationships with defendants. Indeed, the attorney-client relationship is a privileged one, and many aspects of it are afforded formal protection. This underscores the importance of the relationship: the criminal defense attorney often is the defendant's lone ally, providing support and advice and guiding the defendant through the criminal justice process (Lyon, 2012). A breakdown in this relationship leaves a defendant isolated, facing loss of liberty or worse without any support on which to lean. Some have argued that racial bias is likely to precipitate such a breakdown (Pearce, 2005).

One aspect of the attorney-client relationship may be particularly important: the attorney's initial evaluations of his or her client. Research in social psychology shows that people judge the traits and characteristics of others nearly immediately. Traits such as aggressiveness, likeability, trustworthiness, attractiveness, and competence of strangers are judged within 100 milliseconds of their first encounters (Willis & Todorov, 2006). These first

impressions can be difficult to overcome. Once people have been judged as cold, mean, or untrustworthy, there is very little that can be done to change these perceptions (Tausch, Kenworthy, & Hewstone, 2007). Thus, the initial evaluation an attorney makes of a prospective client likely stands as an important factor affecting the full arc of the attorney-client relationship.

We conceptualize "client regard" as consisting of cognitive and affective factors conveying attorneys' appraisal of and affinity for their clients. Client regard is a major component of the attorney-client relationship. Substantial bodies of work on the psychology of prejudice have demonstrated how biased perceptions and emotions impact people's actual and intended behaviors (Schütz & Six, 1996). This basic research has spurred investigations into how stigma against clients might occur in applied professional contexts. In particular, the research has been extensive in the medical domain, with prior work, for example, finding regard to be lower for patients with substance use disorders compared to other illnesses (Meltzer et al., 2013) and for LGBT patients (Madhan, Gayathri, Garhnayak, & Naik, 2012). Plackowski, Hull, and Hull (2014) found client regard to be correlated with behavioral intentions, including interpersonal helping and advocacy. Following this line of research, we are considering client regard[3] as a component of the attorney-client relationship that likely affects attorney-client interactions.

Assessing Racially Biased Attitudes

Scholars have codified two conceptually distinct forms of measuring racial bias: explicit and implicit (Fazio & Olson, 2003; Greenwald & Krieger, 2006). Explicit measures

[3] I would like to note a specific tension here: when we think of criminal defense, we think of attorneys representing the accused, regardless of whether they are serial killers or Neo-Nazis or whomever. From a psychological perspective, we can see how client regard would matter, but from a legal one, should client regard even enter into the discussion? I argue that it should, as the psychological forces at play may carry over into the quality of the legal representation. In an adversarial system, the defendant's attorney is the defendant's sole representative and has a duty to wholeheartedly pursue the defendant's rights. If client regard impacts the fulfillment of this duty, then it has legal significance, and its valence and relative strength may matter, especially if systematically more negative for defendants from certain racial groups.

conceptualize racial attitudes as propositions or declarative beliefs that people can articulate, and they typically assess bias by directly asking people about their attitudes. Explicit biases are understood as attitudes that people formulate, process, and hold within their conscious awareness. On consideration, people can alter the content (Gawronski & Bodenhausen, 2006) or expression (Nier, 2005) of such attitudes. As a result, explicit racial attitudes are particularly vulnerable to social desirability biases, such that, if research participants suspect their negative views are counter-normative or unflattering, they may profess more positive explicit attitudes than they actually possess (Krumpal, 2013). The present research, in gathering information regarding sensitive explicit racial biases from practicing attorneys, is thus a situation in which there is particularly high demand for participants to avoid reporting racial bias.

In contrast, implicit bias measures assess the automatic cognitive associations individuals have for a social group. Racial biases measured at the implicit level are less readily available to introspection (Greenwald & Banaji, 1995) and less subject to conscious control (e.g., see Greenwald, McGhee, & Schwartz, 1998). There is ongoing debate about the most appropriate conceptualization of implicit biases; for example, whether they are a form of personal prejudice, reflect knowledge of cultural stereotypes, or are something else (Arkes & Tetlock, 2004; Fazio, Jackson, Dunton, & Williams, 1995; Payne, Vuletich, & Lundberg, 2017). Regardless of such theoretical debates, implicit bias measures consist of a set of indirect instruments that are considerably less subject to the sort of social desirability concerns that attach to explicit measures (Banse, Seise, & Zerbes, 2001; Nier, 2005).

Explicit and implicit racial biases further differ in that self-reported explicit racial biases in the population have decreased significantly over time (Schuman, Steeh, Bobo, & Krysan, 1997), while implicit racial biases have been more resistant to change, decreasing at much lower

rates (Charlesworth & Banaji, 2019; Nosek et al., 2007). This relative persistence of implicit bias made us want to anchor our focus on the role it may play in criminal defense. Given evidence of implicit bias for White individuals over Black individuals in the general population and given Eisenberg and Johnson's (2004) finding that White capital defense attorneys displayed implicit White preference bias, our first hypothesis was that criminal defense attorneys would, in general, hold implicit White preference bias.

Implicit Racial Bias and Behavior

There is active debate concerning the relationship between implicit bias and more overt attitudes and behavior. In a meta-analysis, Greenwald, Poehlman, Uhlmann, and Banaji (2009) found relatively robust predictive value for the IAT in relation to racial and ethnic discrimination, especially when coupled with self-report measures. In their own meta-analysis, which was based on a sample that included 126 studies, Hofmann, Gawronski, Gschwender, Le, and Schmitt (2005) found a positive mean population correlation of .24 between explicit attitudes and IAT results. However, in another meta-analysis, Oswald, Mitchell, Blanton, Jaccard, and Tetlock (2015) concluded that the IAT is a poor predictor of racial and ethnic discrimination (but see McConnell & Leibold, 2009).

While a variety of explanations have been proposed for these heterogeneous findings, one prime candidate is that the predictive capacity of implicit attitudes is moderated by context and outcome characteristics (Dasgupta, 2004). Thus, for the present study, research focused on the legal context is especially important, and this body of research has consistently established that implicit bias can have an impactful role on emotional and behavioral responses. Fazio and Hilden (2001) demonstrated that, after participants were led to misperceive a Black police officer as a criminal, those with greater implicit bias felt less guilty about the misperception. Levinson,

Cai, and Young (2010) used two IATs, one a traditional IAT and the other an IAT created specifically to examine implicit connections with the presumption of innocence, to show that IAT scores predicted the way in which participants evaluated ambiguous trial evidence. Indeed, much theoretical and empirical research suggests that implicit biases predict spontaneous behaviors and judgments in ambiguous, cognitively demanding situations (Gawronski & De Houwer, 2014; Olson & Fazio, 2009), that is, the type of situations that commonly arise within attorneys' work environments.

Further, implicit racial biases are likely to affect social interactions and interpersonal relationships. Asendorpf, Banse, and Mucke (2002) found that implicit biases correlated with spontaneous behaviors such as shyness, and Jacoby-Senghor, Sinclair, and Shelton (2016) found them to relate to anxiety during interracial interactions. Dovidio, Kawakami, and Gaertner (2002) attempted to clarify the components of the relationship between deliberate behaviors (operationalized as verbal responses) and spontaneous ones (operationalized as perceived nonverbal friendliness). They found that implicit racial bias predicted biased nonverbal behavior but not verbal behavior, and that this nonverbal behavior drove the extent to which participants were perceived as being friendly. Towles-Schwin and Fazio (2003) found that implicit racial bias correlated with expected discomfort when interacting with a Black individual, expectations which can ultimately lead to avoiding such interactions (Plant, Butz, & Tartakovsky, 2008). Together, these findings suggest that implicit racial biases impact interracial interactions in numerous ways: increasing discomfort, decreasing quality, and leading to avoidance of such interactions altogether. This confluence of outcomes does not bode well for Black defendants seeking legal representation or working with legal representatives.

Penner and colleagues (2016) conducted illustrative work on professional relationships (specifically, the physician-patient relationship), finding that implicit bias was negatively associated with physicians' regard for their Black patients. Greater implicit bias resulted in more problematic interpersonal behaviors on a number of dimensions, including shorter conversation length and diminished patient recall of relayed information. Like physicians, and in contrast to other actors in the criminal justice system, defense attorneys labor on behalf of their clients, ideally working in close contact and with common goals. In this chapter, we adapted Penner and colleagues' (2016) research from the medical realm to the legal realm; and, combining it with the basic research outlined above, we were led to our second hypothesis: that criminal defense attorneys' implicit bias would correlate with their interpersonal regard for their clients, such that greater White preference would correlate with worse regard for their Black clients.

Racial Disparities and Drug Stigma

Racial disparities in conviction and incarceration are especially extreme in the context of drug offenses (Alexander, 2010; Quigley, 2012). In some U.S. states, Black men have been convicted on drug charges at rates dozens of times greater than White men (Human Rights Watch, 2000). In 2016, there were more than 81,000 individuals in federal prisons for drug offenses, and 37.9% of them were Black, while 21.6% were White (Carson, 2018). Moreover, a number of reports have shown that the percentage of Black individuals imprisoned for drug crimes is not proportional to arrest rates (Chin, 2013). In other words, once arrested for a drug crime, Black individuals are more likely to go to prison. These disparities in drug convictions exist in spite of the fact that rates of using and selling drugs are similar for White and Black individuals (Substance Abuse and Mental Health Services Administration, 2017).

In the present research, we extended our hypothesis regarding the relationship of criminal defense racial biases to include the impact of drug stigma, such that we predicted that the consequences of criminal defense bias would be especially invidious in the arena of drug offenses. Racial bias, especially in the criminal context, has been well-documented. Studies have found evidence of cultural stereotypes linking African Americans with weapons (Payne, 2006), violence (Payne, 2001), dangerousness (Trawalter, Todd, Baird, & Richeson, 2008), and criminality (Eberhardt, Goff, Purdie, & Davies, 2004). Stigma against drug users also has been well-documented (see Avery & Avery, 2019). Substance use disorders (SUDs), which are considered mental illnesses by all major medical authorities, including the authors of *DSM-5* (American Psychiatric Association, 2013), are viewed more harshly than other forms of mental illness (Rasinski, Woll, & Cook, 2005), and those who use drugs are typically perceived as low in both warmth and competence, perceptions that lead to contempt and social avoidance (Pachankis et al., 2017). As a consequence, the combined effect of racial bias in the unique context of drug offenses places Black individuals in the crosshairs of two emotionally charged factors.

Thus, Black drug offenders may face a confluence of bias and stigma (see Alexander, 2010). Rush (1998) found that participants assigned greater blame to Black individuals who developed drug dependence, as compared to White individuals who did the same. In a qualitative study of 10 Black substance users, Scott and Wahl (2011) found that Black participants perceived that their substance use problems were viewed less favorably than the substance use problems of White individuals. Still others have identified disparate treatment in public policy, arguing that, at the onset of the current opioid epidemic, minority substance users were

considered irresponsible and criminal while White substance users were viewed as suffering from mental disorders and in need of treatment (Netherland & Hansen, 2016).

This foundation yielded our third hypothesis: that, when the clients were Black, defense attorneys' implicit bias would correlate most strongly with their regard toward drug crime (vs. non-drug crime) clients, such that greater implicit White preference would correlate with worse explicit regard for Black drug clients.

In total, the present study builds on literatures documenting the pervasiveness of White preference implicit bias, the relationship between implicit bias and the quality of interracial interactions, and the presence of racialized drug stigma to provide greater insight into the attitudes of criminal defense attorneys towards their clients. Specifically, we tested participants on a race Implicit Association Test, and we presented them with mock clients, about whom we manipulated two variables: race and type of crime committed. Our hypotheses were threefold. First, we hypothesized that criminal defense attorneys would, in general, hold significant White preference implicit bias. Second, we hypothesized that their implicit bias would correlate with their interpersonal regard for Black defendants, such that greater implicit White preference would correlate with worse regard for Black defendants. Third, we hypothesized that implicit racial bias would additionally manifest as drug stigma: implicit White preference would correlate most strongly with worse regard for Black drug (vs. property) defendants.

Method

Participants

In collaboration with state and local bar associations, public defense associations, public defense offices, and other U.S. criminal defense associations, we recruited a unique participant pool of practicing criminal defense attorneys. In the introductory email, the attorneys were told

that this was a survey on attorney work attitudes. Informed consent was obtained from all participants. Participants were excluded if they were not licensed attorneys currently practicing criminal defense. Our final sample included everyone who completed the survey within the approximately two-month timeframe established before distribution. This procedure yielded 327 criminal defense attorneys who practiced in 43 U.S. states. One hundred and fifty-four participants were private criminal defense attorneys, 152 were public defenders, and 21 indicated that they were currently practicing criminal law but did not neatly fit into either of the two categories. Mean years (± SD) in practice was 15.2 ± 10.9. The mean age (± SD) of the participants was 43.9 ± 11.8. The sample was 51.7% male and 48.3% female. The racial breakdown was 89.2% White/Caucasian, 1.2% Black/African American, 4.0% Hispanic/Latino/Latinx, 2.5% Asian/Pacific Islander, .3% Native American, and the rest selected "other." These demographics are consistent with the race and gender demographics of attorneys in the United States (see American Bar Association, 2019), which show a racial breakdown of 85% White/Caucasian, 5% Black/African American, 5% Hispanic/Latino/Latinx, 2% Asian/Pacific Islander, and 1% Native American.

Measures

Implicit Association Test

Implicit bias was measured using a race Implicit Association Test (see Greenwald, McGhee, & Schwartz, 1998) administered on Inquisit software (Millisecond Software, Seattle, WA). We administered the Implicit Association Test in accordance with standard protocol (see Greenwald, Nosek, & Banaji, 2003; Nosek, Greenwald, & Banaji, 2005). Participants were asked to pair face and object category pairs. For example, they were instructed to pair positive words with White faces and negative words with Black faces (compatible condition). The positive

words ("good" category) were marvelous, superb, pleasure, beautiful, joyful, glorious, lovely, and wonderful, and the negative words ("bad" category) were tragic, horrible, agony, painful, terrible, awful, humiliate, and nasty. The faces used are commonly used in the race IAT and can be found on the Project Implicit website (see Nosek et al., 2007). Participants completed multiple trials, including ones in which these pairings changed, i.e., they were asked to pair negative words with White faces and positive words with Black faces (the incompatible condition). Specifically, after a pretrial pause of 250 ms, stimuli were presented until the correct response was given. Participants completed 24 trials of block compatible practice in which the responses were not counted. Then they completed 72 trials under the block compatible condition. Next, they completed 24 trials of block incompatible practice in which the responses were not counted. Then they completed 72 trials under the block incompatible condition. The order of the compatible and incompatible conditions was counterbalanced. The computer software that gathered the data (Inquisit) measured the number of milliseconds it took for participants to respond to each task and calculated a "D-score" by evaluating the difference between incompatible and compatible mean reaction times divided by the pooled standard deviation of reaction times on compatible and incompatible blocks (see Greenwald, Nosek, & Banaji, 2003). The D-score reflected whether participants held implicit associations between the attitude object (good/bad) and the dimension (Black/White) tested. Negative scores indicated anti-White bias and positive scores indicated anti-Black bias,[4] with distance from zero reflecting the magnitude of the bias. Participants' IAT scores were excluded if they had a response latency of < 300 ms on

[4] Conceptually, for this White-Black version of the IAT, negative scores indicate anti-White bias and pro-Black bias, since participants who achieve such scores are better (in terms of accuracy and speed) at pairing White faces with negative terms and Black faces with positive terms. Positive scores indicate pro-White bias and anti-Black bias. For ease of communication, and also for technical accuracy, in this chapter we refer to pro-White/anti-Black bias as "White preference" and pro-Black/anti-White bias as "Black preference."

more than 10% of the response trials, an attentional benchmark established by Greenwald, Nosek, and Banaji (2003). We used this comparative IAT because attorneys, and especially criminal defense attorneys, have limited resources, including time, and must make decisions about their clients in comparison to other clients. For example, given a full caseload, an attorney must decide which clients will receive relatively more attention, care, and support.

Client Regard Scale

We adapted the Medical Condition Regard Scale (MCRS) for use in the current study. The MCRS was designed to capture physicians' conscious biases and emotions towards patients (Christison, Haviland, & Riggs, 2002). It is a valid and reliable 11-item instrument graded on six-point Likert scales (Christison, Haviland, & Riggs, 2002). To create the Client Regard Scale (CRS), we altered the items to reflect the attorney-client relationship instead of the physician-patient relationship. We scored it so that higher scores indicated lower regard (i.e., higher scores indicated worse regard for the client, with 6 being worst and 1 being best). Example items included "I prefer not to work with clients like this," "I enjoy giving extra time to clients like this" (reverse scored), "I feel especially compassionate toward clients like this" (reverse scored) "There is little I can do to help clients like this," and "Representing clients like this is a waste of resources" (Cronbach's $\alpha = .89$).

Predicted Recidivism

Predicted recidivism was measured with a single item—"In your opinion, how likely do you think it is that this individual will be arrested (for any crime) at some point in the future?"—graded on a six-point Likert scale.

Demographics Questions

Participants completed a demographics questionnaire reporting information such as race, gender, age, political party, geographic region, and type of law they were practicing.

Design and Procedure

We implemented a mixed design, with client race (White, Black, race-unspecified) as the between-subjects variable and crime type (drug offense, property offense) manipulated within-subjects. In the procedure, participants were told to imagine that they had taken on new clients, and they viewed two "crime profiles" that conveyed information about fictitious defendants and their alleged crimes.

Client demographic information was delivered through a combination of a picture and a name that were either stereotypically White, Black, or "neutral" (wherein no pictures, names, or other race indicative information was included). We pretested pictures and names to be comparable according to perceived age, attractiveness, race prototypicality, and socioeconomic status (SES). We standardized for these particular dimensions because of previous research indicating differential treatment of legal parties based on age (Zatz, 2000), attractiveness (DeSantis & Kayson, 1997), race prototypicality (Blair, Judd, & Chapleau, 2004), and SES (D'Alessio & Stolzenberg, 1993). All faces were drawn from the Chicago Face Database (Ma, Correll, & Wittenbrink, 2015). Participants viewed two different individuals within the same demographic category across the two crime profiles they considered. The defendant pictures were counterbalanced across crimes. Client race was kept consistent within participants in the hopes of minimizing participant awareness that race was of experimental interest.

Participants considered both drug and property offenses. In the drug condition, participants were told the client had just been arrested and charged with unlawful possession of a controlled substance, a felony, and the client's record revealed that, a few years ago, the client

had been convicted of a similar drug offense. In the property condition, participants were told the client had just been arrested and charged with felony larceny, and the client's record revealed that, a few years ago, the client had been convicted of a similar property offense. The descriptions of these crimes were tailored so that they were equivalent in severity according to the U.S. Sentencing Commission's 2016 Sentencing Guidelines Manual. Each participant saw one property crime and one drug crime, with the order of presentation counterbalanced so as to prevent order effects.

After viewing each crime profile, participants indicated their regard for the client using the CRS as well their beliefs regarding the likelihood that the client would recidivate. Finally, participants completed a race IAT. Participants then provided demographics information.

For their participation, participants were entered into a lottery to receive one of three $100 gift certificates. The research protocol was approved by the Princeton University Institutional Review Board.

Results

Implicit Bias

Our first hypothesis was that criminal defense attorneys would show relative White preference implicit bias. Evidence for implicit bias was examined using a one-sample two-tailed t-test of D-scores. D-scores were significantly greater than zero (M = .29, SE = .03, 95% CI [.24, .34], $t(326) = 11.41$, $p < .001$, Cohen's d = .63, 95% $CI_{\text{Cohen's } d}$ [.50, .76]), such that participants on average exhibited a pattern of stereotypic trait associations with White and Black faces (Figure 1.1). In other words, participants showed a preference for White individuals (vs. Black). 75.8% of participants showed White preference.

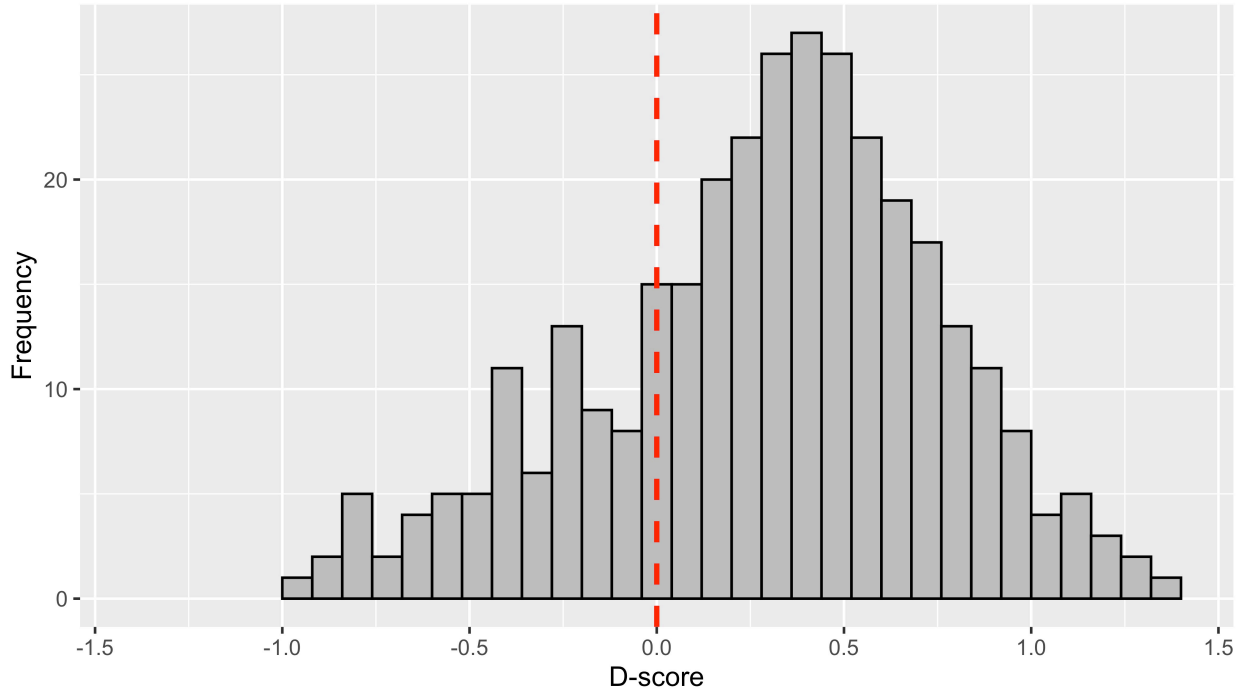


Figure 1.1 Frequency distribution of Implicit Association Test scores (D-scores). Negative D-scores indicate relative Black preference and positive D-scores indicate relative White preference, with distance from zero reflecting the magnitude of the bias. A score of 0.0 (indicated by the dashed line) represents no relative preference.

In an attempt to minimize participants' awareness that we were interested in the effect of defendant race, participants were shown the crime profiles and completed the CRS and recidivism measures before taking the IAT (for a similar design, see Green et al., 2007; see also McConnell & Leibold, 2001). This ordering, of course, left open the possibility of priming effects. We did find slight effects, such that participants who read about Black defendants (M = .39) showed greater White preference than those who read about White defendants (M = .23) or defendants in the race-unspecified condition (M = .24). Nevertheless, one-sample two-tailed t-tests confirmed that the means for Black, White, and race-unspecified all were significantly greater than zero (Black: M = .39, SE = .04, 95% CI [.30, .47], $t(125) = 9.33$, $p < .001$, Cohen's d = .83, 95% $CI_{\text{Cohen's } d}$ [.57, 1.10]; White: M = .23, SE = .04, 95% CI [.15, .31], $t(133) = 5.72$, $p <$

.001, Cohen's d = .49, 95% $CI_{Cohen's\ d}$ [.32, .66]; race-unspecified: M = .24, SE = .05, 95% CI [.14, .35], $t(66) = 4.53$, $p < .001$, Cohen's d = .55, 95% $CI_{Cohen's\ d}$ [.28, .83]), such that all exhibited significant implicit White preference bias.

While a number of participants did not fully answer demographics questions, likely on account of privacy and/or social desirability concerns, we conducted an exploratory analysis of participants' characteristics and what they suggested about implicit White preference bias. When we regressed implicit bias in a linear model including all participant characteristics (race, age, gender, political party, geographic region), we found that participants who were White, older, Republican, and from a southern state[5] were more likely to show greater White preference: race, where 0 is non-White and 1 is White (β = .31, SE = .10; $t(212) = 3.24$, p = .001); age (β = .005, SE = .003; $t(212) = 2.06$, p = .04); political party, where 0 is Republican and 1 is Democrat (β = -.15, SE = .08; $t(212) = -1.98$, p = .049); geographic region, where 0 is non-southern states and 1 is southern states (β = .13, SE = .07; $t(212) = 1.98$, p = .049). Gender was not significant: where 0 is male and 1 is female (β = -.04, SE = .06; $t(212) = -.67$, p = .50).

Implicit Bias and Client Regard

Our second hypothesis was that defense attorneys' implicit bias would correlate with their regard for Black clients, such that greater implicit White preference would correlate with lower regard for Black clients.

Client Regard

Following Penner et al. (2016), we analyzed attorneys' regard for their Black clients on the Client Regard Scale. Due to the fact that each attorney viewed two client profiles and thus

[5] The South was defined as the twelve states that had the most lynchings between 1877 and 1950, which include Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Texas, and Virginia (Tolnay & Beck, 1995). This is the traditional "Deep South."

contributed two observations, we employed a linear mixed-effects model to account for group effects from participants (that is, attorneys). We fit a linear mixed-effects model (mixed-effects regression, by-participant, with random intercept) by maximum likelihood (see Gelman & Hill 2006), and we found that the implicit bias of attorneys correlated with their total CRS scores (where greater CRS scores signified greater attitudinal negativity), $\beta = .28$, $t(124) = 2.80$, $p = .005$ (Figure 1.2), such that greater implicit White preference correlated with worse regard for Black clients.[6]

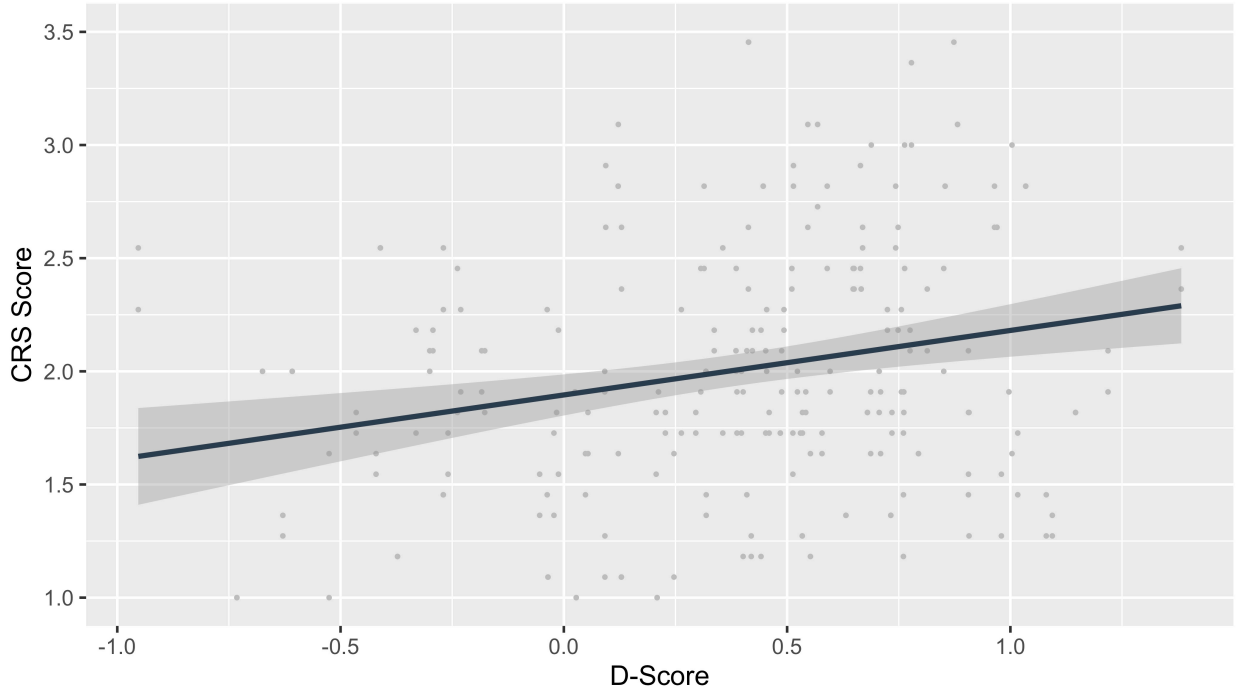


Figure 1.2 Plot of model of regard for Black clients (CRS score, where greater CRS scores indicate greater attitude negativity, i.e., worse client regard) and implicit bias (D-score, where greater D-scores indicate greater White preference). The shaded region is the 95% confidence interval (p = .005).

[6] A post-hoc power analysis showed that we were appropriately powered for detecting a medium effect size, given a significance level of .05.

As exploratory analyses, we repeated this with the other groups. The relationship between IAT D-scores and CRS scores was not found when we considered only White clients ($\beta = .18$, $t(132) = 1.04$, $p = .30$), nor was an effect found when we considered the race-unspecified group ($\beta = .23$, $t(65) = 1.38$, $p = .18$). When all race conditions were included in a single model (again a linear mixed-effects model fit by maximum likelihood), the two-way interaction of race and implicit bias on CRS scores failed to reach significance. This was unsurprising given the impact of the race-unspecified group, which fell at the midpoint across the racial groups, and also, more importantly, given our sample size. To test whether the two-way interaction failing to reach significance may have been due to an insufficient number of participants, we conducted a post-hoc power analysis. Assuming a linear model in which the two-way interaction accounts for 20% of the total variance (5.9%) explained, the power was .522, showing that we were, indeed, underpowered for detecting the two-way interaction, as the minimum sample size required to achieve 0.8 power would be 626 participants.

Predicted Recidivism

Employing a linear mixed-effects model (mixed-effects regression, by-participant, with random intercept) fit by maximum likelihood, we found that, for Black clients, the correlation between the implicit bias of attorneys and their predicted recidivism scores (where greater scores signified greater belief that the client would reoffend) was $\beta = .28$, $t(122) = 1.91$, $p = .06$. No effect was found when we performed the analysis for the group containing only White clients, $\beta = .10$, $t(132) = .86$, $p = .39$, nor when we performed the analysis for the group containing only race-neutral clients, $\beta = .10$, $t(65) = .55$, $p = .58$. When all race conditions were included in a single model (again a linear mixed-effects model fit by maximum likelihood), the two-way

interactions of race (White, Black, unspecified) and implicit bias on predicted recidivism failed to reach significance.

Implicit Bias and Client Regard for Black Drug Offenders

Our third hypothesis was that, while implicit White preference would be positively correlated with more negative regard for Black clients, the effect would be largest when the client had committed a drug offense (vs. non-drug offense).

Client Regard

We employed Ordinary Least Squares (OLS) linear regression models and found that, as predicted, when the clients were Black drug offenders, the implicit bias of attorneys correlated with their total CRS scores: β = .32, 95% CI [.11, .53], $t(124) = 3.03$, $p = .003$. We found a smaller effect for Black property offenders (β = .25, 95% CI [.03, .47], $t(124) = 2.26$, $p = .03$), but the difference between the effects was not significant. The relationship between IAT and CRS scores was not significant for any of the other groups (White drug offenders: β = .10, 95% CI [-.14, .34], $t(132) = .82$, $p = .41$; White property offenders: β = .15, 95% CI [-.13, .44], $t(132) = 1.08$, $p = .28$; race-unspecified drug offenders: β = .22, 95% CI [-.14, .58], $t(65) = 1.21$, $p = .23$; race-unspecified property offenders: β = .25, 95% CI [-.11, .60], $t(65) = 1.39$, $p = .17$). When all conditions were included in a single linear mixed-effects model fit by maximum likelihood, the three-way interaction of race (White, Black, unspecified), crime-type (drug or property), and implicit bias on CRS scores failed to reach significance.

Predicted Recidivism

As with client regard, the individual models for predicted recidivism showed results in line with our predictions: predicted recidivism was correlated with IAT scores only for Black drug offenders (β = .32, 95% CI [.00, .63], $t(122) = 1.99$, $p = .049$). When all conditions were

included in a single linear mixed-effects model fit by maximum likelihood, the three-way interaction of race (White, Black, unspecified), crime-type (drug or property), and implicit bias on predicted recidivism scores was significant: $\beta = .22$, $SE = .10$, $p = .04$.

Discussion

Scholars have long argued that criminal defense attorneys are likely to harbor racial implicit bias, but few data supporting this thesis have been gathered (see Kang et al., 2012; Richardson & Goff, 2013). The present study, which surveyed a sample of U.S. criminal defense attorneys from 43 U.S. states, gathered just such data—data that may form a foundation for thinking about bias and disparate outcomes in criminal defense. Specifically, our findings showed that, on average, criminal defense attorneys harbored implicit White preference bias.

Seventy-six percent of our participants showed White preference implicit bias. For comparison, Banaji, Bazerman, and Chugh (2003) considered data gathered from Project Implicit and found that 75% of participants showed implicit White preference bias. Our criminal defense attorney participants had an average D-score of .29, which is in the slight-to-moderate bias range (see Xu, Lofaro, Nosek, & Greenwald, 2018).[7] Given that the only previous collection of criminal defense attorney IAT scores comes from the study by Eisenberg and Johnson (2004) that contained a niche subset of defense attorneys (capital attorneys who either worked in Georgia at the trial level or worked on federal habeas corpus proceedings),[8] our findings are an important addition to the literature on the magnitude of bias in criminal defense.

Implicit attitudes may affect interpersonal interactions, which are paramount in criminal defense. As discussed above, Penner et al. (2016) found that greater IAT scores were associated

[7] As a further comparison, we analyzed data from individuals who, from 2006 and 2015, took a race IAT on Project Implicit's platform, and D-scores were .32.
[8] Eisenberg and Johnson (2004) also used an older, pen-and-paper version of the IAT.

with more negative interpersonal interactions between physicians and Black patients. Mendes and colleagues found that implicit bias was linked to greater stress arousal during interracial interactions (Mendes, Gray, Mendoza-Denton, Major, & Epel, 2007). Research has consistently shown that implicit White preference attitudes among interaction partners can produce changes in the behavior of Black individuals, changes that can be interpreted as confirmatory of the validity of the bias held (Word, Zanna, & Cooper, 1974). In other words, individuals' implicit negative associations may shape interracial interactions in ways that lead to negative self-fulfilling prophecies.

Whatever effect implicit bias has at the level of the attorney-client (the one-on-one level), the aggregate impact of such bias is likely to be considerable (Payne, Vuletich, & Lundberg, 2017; see also Greenwald, Banaji, & Nosek, 2015). Disparate outcomes stemming from implicit bias tend to be largest when examined in aggregate, such as at the county or city level (see Riddle & Sinclair, 2019), or at the level of organizations, such as the criminal justice and public defense organizations in which many criminal defense attorneys work (Payne, Vuletich, & Lundberg, 2017). Thus, the sum total effects are likely quite large, suggesting a need for effective interventions.

One's Own Team

Anyone who confronts the criminal justice system needs to navigate a confusing array of situations and actors. It is understandable for defendants to view police and prosecutors as adversaries in the process, just as it reasonable for them to harbor concerns about judges and jurors. However, the accused's partner in the system is her or his attorney.

Although the implicit bias-client regard nexus was difficult to unpack in the present research, we did find suggestions that implicit bias mattered in this way. When the defendants

were Black, implicit bias and client regard were correlated, such that greater implicit White preference was associated with worse regard for Black defendants. This is an important nexus to explore further, since a key aspect of criminal defense is the interpersonal relationship that develops between the attorney and his or her client. A strong attorney-client relationship is essential for ensuring that defendants are confident in the system, well-informed about their cases, and poised to make thoughtful legal decisions. Imagine that you have been assigned a partner for an important project. Now also imagine that, for some unknown reason, the partner does not like you. How will this affect your performance? How will it affect your emotional state? How will it affect your decision making? This is the criminal defense attorney-client relationship in which many minority defendants find themselves.

Finally, given the importance of eliminating racialized attitudes in criminal defense, a case can be made for interventions that begin as early as possible, such as during law school. Law students should be aware of implicit racial biases and their relationship to explicit attitudes, and they should be aware of the potential confluence of racial bias and addiction stigma, especially given the fact that clinics and internships for law students often involve providing legal support to minority, underserved, and vulnerable populations. That said, bias is an insidious problem, and interventions focused on correcting for bias have had only limited success (Forscher et al., 2019; Hetey & Eberhardt, 2018; Payne et al., 2017). While this chapter investigates the problem and is not focused on how it might be fixed, we do caution that creative interventions, perhaps the use of technology as an anchoring device, might be needed if progress is to be made (Avery & Cooper, 2020b; Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018).

Limitations and Future Directions

First, and as discussed above, we were underpowered to detect some of the more nuanced effects, such as interactions across client race. The special nature of the participant pool (i.e., criminal defense attorneys) makes it challenging to recruit very large numbers. Nonetheless, if, in future work, researchers are able to attract the requisite number of criminal defense participants, then such interactions might be more analyzed and unpacked more robustly.

Second, attitudes about future criminality mirrored our client regard results. When the defendants were Black drug offenders, predicted recidivism was correlated with IAT scores. Moreover, when all conditions were included in a single linear mixed-effects model fit by maximum likelihood, the three-way interaction of race (White, Black, unspecified), crime-type (drug or property), and implicit bias on predicted recidivism scores was significant. Thus, future work might explore the extent to which beliefs about defendants' states and traits might influence attorney behavior. For instance, we might hypothesize that, when the defendant is Black and has committed a drug offense, attorneys view this as a character flaw (a trait), as a defect that will lead to future crimes, rather than as a misstep (a state). Further empirical work could shed light on this.

Third, one potential next step in this line of research is clear. Given that we have established implicit bias and taken steps towards understanding its importance for the attorney-client relationship, future work might evaluate additional outcomes on both sides. For attorneys, when there is low regard for a client, does this impact strategic decision making, including judgment related to plea deals? For clients, when there is low regard from one's attorney, does this affect client well-being, decision making, and trust in the criminal justice system? While such outcomes would be challenging to gather, and may require durative field work, the potential yield for psychological understanding of bias in the law is large.

Conclusion

The present research provided evidence that criminal defense attorneys held significant implicit White preference bias. When the defendants were Black, implicit bias and client regard were inversely correlated, such that greater implicit White preference was associated with worse regard for Black defendants. In a criminal case, the attorney-client relationship is paramount, and interpersonal regard is a key element of any such relationship. Our results suggest that this relationship may be impacted by bias, thus increasing the difficulty that Black defendants face. In the ongoing scholarly endeavor to explain racially disparate rates of incarceration in the United States, it is becoming clear that the psychology of criminal defense attorneys could help to provide an answer. In Chapter 2, I explore this possibility further by delving deeper into criminal defense decision making.

Note

The research presented in this chapter was conducted by Joel Cooper, Jordan Starck, Jonathan D. Avery, Yiqiao Zhong, and the dissertation author.

Chapter 2

Implicit Bias, Fear of Being Labeled Biased, and Criminal Defense Attorneys' Motivation to Correct for Bias in the Law

To delve deeper into the issues explored in Chapter 1, I partnered with three statewide criminal defense associations (Texas, Virginia, Washington State) and one national organization (the National Association of Criminal Defense Lawyers), which distributed a new survey via their listservs. The final population included 183 actively practicing criminal defense attorneys, with a majority (greater than 80%) practicing in Texas, Virginia, or Washington.

The aim was to strengthen our grasp on two threads of thought. One, to what extent is there racial bias in criminal defense? In short, I wanted to replicate and shine additional light on the results presented in Chapter 1. Two, if the path to racial equality in the law requires committed effort by criminal defense attorneys, what might impact whether or not defense attorneys embrace this challenge? This second point stems from my reading of U.S. legal history, especially in the context of race and criminal defense. Although many legal protections were originally created for the express purpose of protecting Black defendants, since the 1960s the Supreme Court has allowed race to recede into the background of its attendant jurisdiction (Ossei-Owusu, 2019). However, racial disparities are a significant feature of the criminal justice system, and scholars are increasingly becoming aware that, unless racial equality is an organizing principle for criminal defense firms and organizations, criminal defense likely will continue to have a worsening effect on racial disparities in legal outcomes (Yaroshefsky, 2016). With this understanding, it was important that we delve deeper into the qualitative aspects of defense work to explore how and when bias might arise—and how and when there might be motivation to lessen both its presence and impact.

Method

The survey was more qualitative than that presented in Chapter 1, and it had four types of questions. First, there were questions dealing with racial bias: whether the attorney him or herself showed bias that favored a particular racial group. This part of the survey included an Implicit Association Test, race explicit attitude thermometers, and the Symbolic Racism measure. For the Implicit Association Test, we used a race version (see Greenwald, McGhee, & Schwartz, 1998) that we administered online using the method developed by Carpenter and colleagues (2018). In addition, we followed the standard psychological protocol for using the test (see Greenwald, Nosek, & Banaji, 2003; Nosek, Greenwald, & Banaji, 2005). Participants were asked to pair face and object category pairs. For example, they were instructed to pair positive words with White faces and negative words with Black faces (compatible condition). The positive words ("good" category) were marvelous, superb, pleasure, beautiful, joyful, glorious, lovely, and wonderful, and the negative words ("bad" category) were tragic, horrible, agony, painful, terrible, awful, humiliate, and nasty. The faces used are commonly used in the race IAT and can be found on the Project Implicit website (see Nosek et al., 2007).

For symbolic racism, we used Henry and Sears's (2002) scale, which was developed to measure endorsement of the following ideas: racism is no longer a threat to Black individuals' prospects for good lives, Black individuals are disadvantaged because they are unwilling to take responsibility for their lives, and thus most proposed ways of helping Black individuals are not justified (Sears, van Laar, Carrillo, & Kosterman, 1997).

Second, there were questions dealing with perceived racial bias in the law. In particular, participants were asked a series of questions regarding the extent to which they believed bias

impacts criminal defense decision making. Third, there were questions dealing with an increasingly important issue: how often were the attorneys accused of bias in their professional capacity, and to what extent did they worry about such accusations? Fourth, we asked participants to write a response to the question, "If you were tasked with eliminating racial bias in the criminal justice system, what would you do?" We then tallied the number of characters written in response to the question, using this as a proxy for effort exerted. Please see the Appendix for a full description of these measures. In addition, the measures are discussed in detail, along with the results, in the results section of this chapter.

Lastly, participants completed a demographics questionnaire that included information such as race, gender, age, years practicing, geographic region, and type of law they were practicing. For their participation, participants were entered into a lottery to receive a $150 gift certificate. The research protocol was approved by the Princeton University Institutional Review Board.

The survey was distributed by email/listserv to members of the criminal defense associations for Texas, Virginia, and Washington State, as well as the National Association of Criminal Defense Lawyers. Participants were excluded if they were not licensed attorneys currently practicing criminal defense. This procedure yielded 183 criminal defense attorneys who practiced in 23 U.S. states, with a majority (greater than 80%) practicing in Texas, Virginia, or Washington. Twenty-five percent were full-time public defenders, and the rest were private criminal defense attorneys, some of whom also represented indigent defendants. Mean years (± SD) in practice was 22 ± 13. The mean age (± SD) of the participants was 52 ± 13. The population was 62% male and 37% female, with remainder selecting transgender or "other." The racial breakdown was 77% White/Caucasian, 4% Black/African American, 12%

Hispanic/Latino/Latinx, 1% Asian/Pacific Islander, 1% Native American, 1% Middle Eastern/Arab, 2% Multiracial, and the rest selected "other" or preferred not to disclose.

Results

Implicit Bias

Our primary hypothesis was that criminal defense attorneys would show relative White preference implicit bias. Evidence for implicit bias was examined using a one-sample two-tailed t-test of D-scores. D-scores were significantly greater than zero (M = .35, SE = .03, 95% CI [.29, .42], $t(182) = 10.32$, $p < .001$, Cohen's d = .76), such that participants on average exhibited a pattern of stereotypic trait associations with White and Black faces (Figure 2.1). In other words, participants showed a preference for White individuals (vs. Black). Seventy-six percent of participants showed White preference.

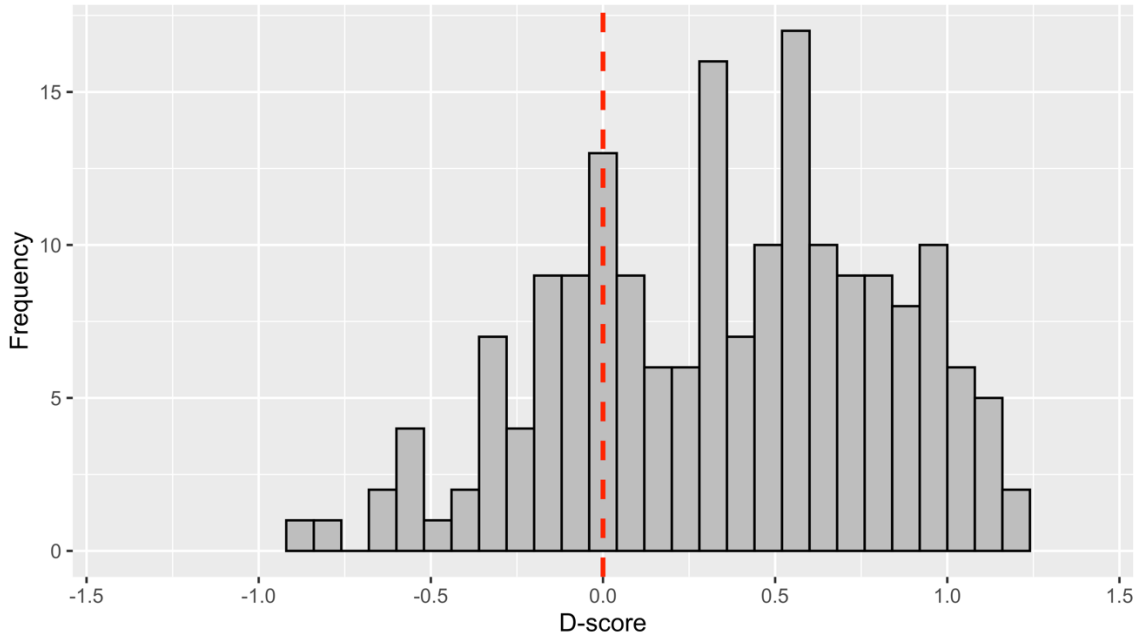


Figure 2.1 Frequency distribution of Implicit Association Test scores (D-scores). Negative D-scores indicate relative Black preference and positive D-scores indicate relative White

preference, with distance from zero reflecting the magnitude of the bias. A score of 0.0 (indicated by the dashed line) represents no relative preference.

When all of the demographic variables were included in a single linear model (see Bruce, Bruce, & Gedeck, 2017; James, Witten, Hastie, & Tibshirani, 2014) with IAT D-scores as the dependent variable, the only significant predictor was age, such that, as the age of the attorneys increased, they showed slightly increasing White preference implicit bias (very slight: $\beta = .01$, 95% CI [.00, .01]). In short, while there were some trends observed, such as greater White preference implicit bias by White (vs. non-White) participants and greater White preference implicit bias by private criminal defense attorneys (vs. public defenders), these trends were not significant.

Effort on Behalf of Racial Justice in the Law

Participants were asked to write a response to the question, "If you were tasked with eliminating racial bias in the criminal justice system, what would you do?" This was our primary dependent variable, outside of the IAT test, as it was designed as a measure of willingness to expend effort for racial justice in the law. As a proxy for effort, we considered the number of characters written in response to the question, and the mean (in number of characters) was this: $M = 273$, $SE = 26$, 95% CI [223, 324].

Additional Measures of Bias

Participants completed the symbolic racism scale and two racial feeling thermometers. For symbolic racism, we used Henry and Sears's (2002) scale, which consists of eight items (see

Appendix). We averaged the items in order to produce a single score. Cronbach's alpha was .90. The results were: $M = 2.11$, $SE = .06$, 95% CI [1.98, 2.24].

Participants also completed two feeling thermometers, presented in random order. The thermometer asked, "How do you feel about White [Black] people?" The response scale ranged from 0 (cold or unfavorable) to 50 (no feeling at all) to 100 (warm or favorable). For the attitudes about White individuals thermometer, the results were: $M = 67.74$, $SE = 1.73$, 95% CI [64.33, 71.15]. For the attitudes about Black individuals thermometer, the results were: $M = 74.63$, $SE = 1.53$, 95% CI [71.61, 77.65]. The difference between the means was significant at $p = .004$.

Fear of Being Labeled Racist

Participants were asked to what extent they worried about being accused of racial bias. The scaled ranged from 1 to 5, with greater scores indicating greater worry (0 = strongly disagree with the statement "I worry about being accused of racial bias"; 3 = neither agree nor disagree; 5 = strongly agree). The results were: $M = 2.60$, $SE = .09$, 95% CI [2.42, 2.77].

They also were asked whether they had ever been accused of racial bias in their professional capacity. If they answered in the affirmative, they then were asked to indicate who, from a list of options, had accused them. Thirty percent indicated that they had been so accused. Of these who had been accused of racial bias, almost all indicated that the accusation had been made by defendants or family members of defendants.

Beliefs about Bias by Criminal Defense Attorneys

Participants were asked various questions about their peers' bias, that is, bias in other criminal defense attorneys. We phrased the questions this way (in terms of *peer* criminal defense

attorneys) because much research has shown that, when asked what others believe, people use others' beliefs as proxies for their own beliefs (Krueger & Clement, 1994; Petsko & Bodenhausen, 2019). With a difficult subject matter, like racial bias, this approach is a useful one, as it may come closer to reflecting actual attitudes and beliefs.

Participants were asked whether they believed racial bias, implicit bias, and explicit bias impact their peers' legal decision making. The scale ranged from 1 (definitely not) to 4 (definitely yes). For racial bias in general, the results were: $M = 2.89$, $SE = .06$, 95% CI [2.76, 3.01]. For implicit bias, the results were: $M = 3.11$, $SE = .06$, 95% CI [2.99, 3.23]. Eighty-three percent of participants believed that implicit bias impacted criminal defense decision making. For explicit bias, the results were: $M = 2.44$, $SE = .09$, 95% CI [2.32, 2.56]. Forty-six percent believed that explicit bias impacted criminal defense decision making. In plain English, we might conclude that the participants do not strongly believe that racial bias in general or explicit racial in particular impacts criminal defense decision making, likely because they believe such attitudes are censured. But they do believe that implicit racial bias is there, operating the background and impacting decision making.

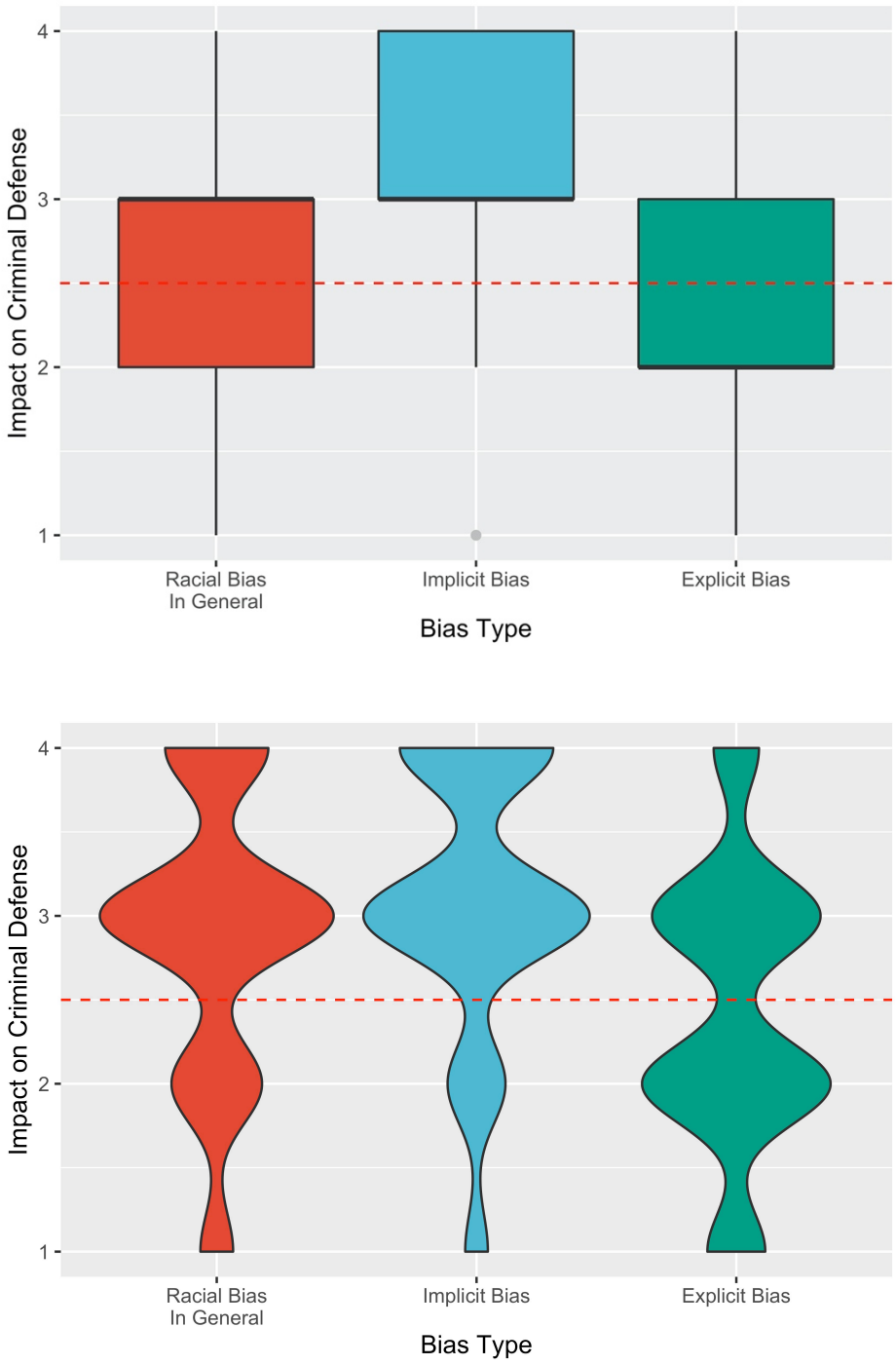


Figure 2.2 Boxplots (top) and violin plots (bottom) of the perceived impact of different types of bias on criminal defense decision making. Scores above the horizontal dotted line indicate beliefs that the bias does have an impact on criminal defense, while those below the line indicate a belief in no impact. The distance from the horizontal line represents the intensity of the belief (i.e., a score of 4 = definitely has an impact, while a score of 1 indicates definitely does not have an impact).

For implicit and explicit bias, if participants indicated a belief that such bias has an impact, they were asked a follow-up question: "What's your best guess as to how often implicit [explicit] racial bias impacts your peers' legal decision making?" The scale ranged from 0 (never) to 100 (all the time). For implicit bias, the results were: $M = 61.63$, $SE = 1.97$, N = 151, 95% CI [57.74, 65.52]. For explicit bias, the results were: $M = 47.87$, $SE = 2.70$, N = 85, 95% CI [42.50, 53.24].

Interventions

Lastly, we asked our participants about interventions. If the goal is that every defendant will be treated fairly, what is important? We provided six options, which related to training (implicit and explicit bias trainings), workloads (lighter dockets), support (more staff), freedom from bias claims (assurance that they can perform their work without being accused of bias; referred to as "bias immunity" in Table 2.1), remuneration (better compensation), and knowledge (information or statistical analyses that can help guide legal decision making). The scale contained the following options: 1 = not at all important, 2 = slightly important, 3 = moderately important, 4 = very important, 5 = extremely important.

Table 2.1 Perceived importance of interventions for decreasing bias in criminal defense, where 1 is not at all important and 5 is extremely important.

Intervention	Mean	SD	SE	95% CI
Training	3.77	1.16	0.09	[3.60, 3.94]
Workload	3.76	1.17	0.09	[3.59, 3.93]
Support	3.48	1.24	0.09	[3.30, 3.66]
Bias Immunity	2.55	1.34	0.10	[2.35, 2.74]
Remuneration	3.65	1.21	0.09	[3.47, 3.83]

Knowledge	3.16	1.28	0.10	[2.97, 3.35]

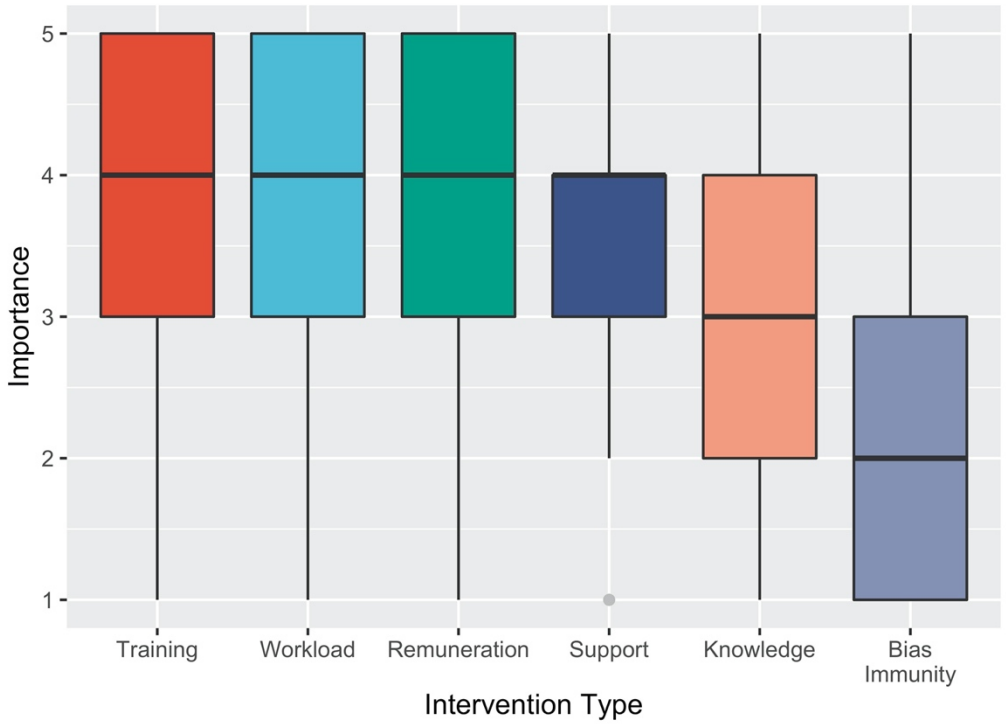


Figure 2.3 Boxplots of the perceived importance of various interventions for decreasing bias in criminal defense, where 1 is not at all important and 5 is extremely important (greater scores indicate greater perceived importance).

The most striking takeaway from the Figure 2.3 is that participants did not want to be let off the hook. They did not want immunity from bias claims, nor did they want more support to help them. Rather, they wanted the core feature of the job addressed—too many cases for too little pay—and they wanted training that could help steer them towards unbiased decision making.

Modeling the Relationship of Attitudes to Effort

To analyze the effects of the various measures discussed above, we included all of the measures relating to race in a single model, with the effort measure (number of words written in formulating a way to combat bias in the law) as the dependent variable.[9] This enabled us to evaluate the relationship between the different bias indicators and effort expended on formulating a solution to racial bias in the law. We employed multiple linear regression estimation using ordinary least squares, and our model was determined via stepwise model selection by AIC (Bruce, Bruce, & Gedeck, 2017; James, Witten, Hastie, & Tibshirani, 2014).

Results of the multiple linear regression indicated that there was a collective significant effect of IAT scores (white preference implicit bias), fear of accusations of bias, belief that bias impacts criminal defense, and age on effort: $F(5, 152) = 3.97$, $p = .002$, $R^2 = .12$. The individual predictors were as follows: IAT scores ($\beta = -.14$, $SE = .07$, 95% CI [-.28, .-.01], $t(152) = -2.05$, $p = .04$); fear of accusations ($\beta = -.10$, SE = .05, 95% CI [-.20, -.01], $t(152) = -2.16$, $p = .03$); belief in bias impact ($\beta = .15$, $SE = .05$, 95% CI [.04, .25], $t(152) = 2.65$, $p = .01$); age ($\beta = .00$, $SE = .00$, 95% CI [.00, .00], $t(152) = 2.21$, $p = .03$). In other words, as white preference implicit bias increased, participants exerted less effort in formulating a solution to bias in the law. Likewise, as participants' fears increased that in their professional capacity they would be accused of bias, they exerted less effort. In contrast, as participants' beliefs that racial bias impacts criminal defense decision making increased, they were more willing to exert effort in formulating a solution to bias in the law. These results are depicted in Figure 2.4.

[9] For a correlation table with all the items included, see Appendix II.

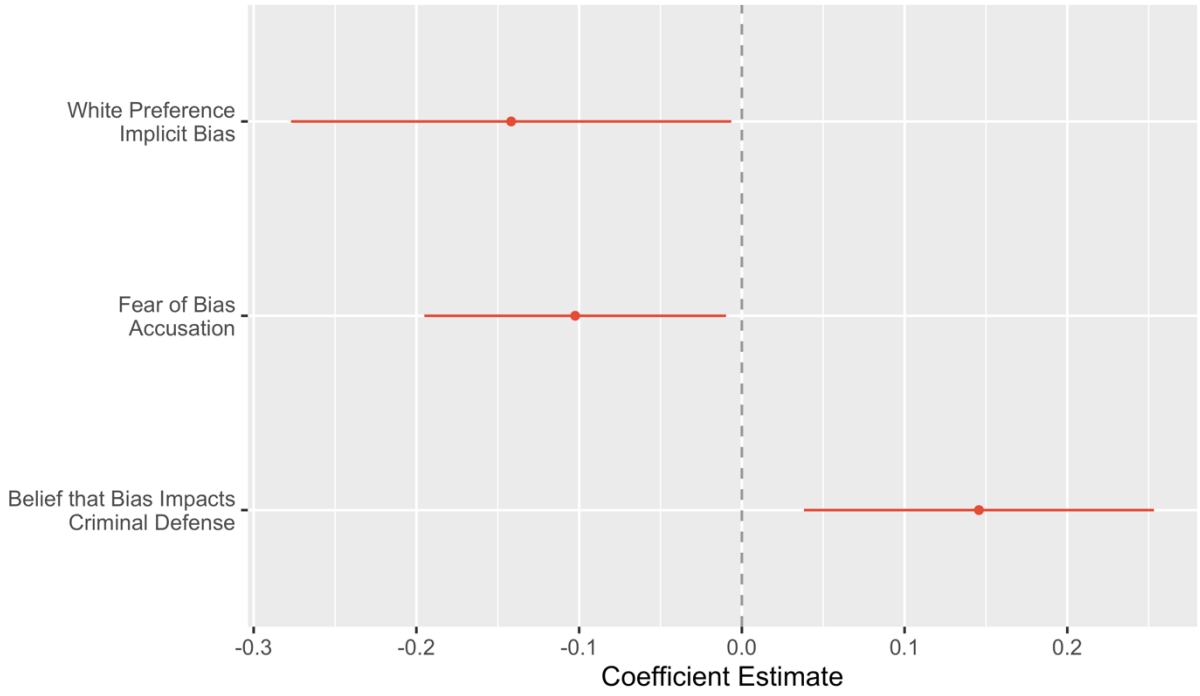


Figure 2.4 Regression coefficients for a measure of criminal defense attorneys' effort expended on a question of racial justice in the law. Age also was a significant factor, but its coefficient was 0.0, so it was not included in this figure.

Discussion

In addition to the effort result, there were three other important results: criminal defense attorneys showed white preference implicit bias, they believed this bias impacted both defense attorneys and judges, and they both worried about and often were accused of bias in their professional capacity. We will consider these results in turn. First, the attorney's implicit bias D-score had a mean of .35, such that participants on average showed a preference for White individuals (vs. Black). Seventy-six percent of participants showed this White preference.

Interestingly, while there were some trends observed, such as greater White preference implicit bias by White (vs. non-White) participants and greater White preference implicit bias by

private criminal defense attorneys (vs. public defenders), these trends were not significant. Indeed, all of these groups—White and non-White attorneys, private and public defense attorneys—showed White preference implicit bias.

Second, on average, the criminal defense attorneys believed that racial bias probably has an impact on criminal defense decision making. In particular, they were most certain that implicit racial bias impacted criminal defense decision making. Those who believed it had an impact also believed that the impact was pretty common: they provided a mean score of 61 for how often did implicit bias impact defense attorneys in their practice, on a 0 (never) to 100 (all the time) scale.

As far as the perceived impact of various factors on judicial sentencing decisions, participants believed that criminal history and prosecutors' recommendations had the most impact. Race came in much lower, with the same average perceived impact as gender: participants indicated a "moderate" role in judicial decision making.

Third, a significant percentage of criminal defense attorneys indicated that they had been accused of racial bias in their professional capacity: a full 30% had been so accused. Of these, almost all indicated that the accusation had been made by their own clients or family members of the clients. However, the fear of being so accused was not overly great. On average, participants neither agreed nor disagreed with the statement, "I worry about being accused of racial bias," and the trend was in the direction of disagreeing with the statement. In other words, they claimed not to worry too much about it.

Conclusion

In addition to replicating the implicit bias results reported in Chapter 1, the study in Chapter 2 paints a more complete picture of bias in criminal defense work. Not only is it

significant and consistently found—both in implicit and explicit measures—but it appears to impact defense attorneys' commitment to addressing racial disparities in the law. Results of the multiple linear regression showed the following significant predictors: implicit bias, worry about being accused of bias, and belief in the impact of bias on criminal defense work. The model thus suggests that, if one wants to know how much effort a defense attorney will expend on racial equality in the law, one should ask three questions. First, is the attorney implicitly biased? Second, does he or she worry about being accused of bias? Third, does the attorney believe that racial bias impacts criminal defense decision making? There was an observed correlation between, one, the attorney's bias attorney; two, how much the attorney worried about being accused of bias; and, three, the outcome measure of effort expended on legal racial equality, such that increases in the first two correlated with a drop in effort.

This is a grave problem. Chief Justice John Roberts wrote that, the right to counsel is "the most precious right a defendant has, because it is his attorney who will fight for the other rights the defendant enjoys" (Kaley v. United States, 134 S. Ct. 1090, 2014, p. 1107). It stands to reason that if the defense attorney is biased against his or her Black clients, and if the attorney is unmotivated to work to ensure that racial disparities do not emerge, then Black clients stand at a significant disadvantage.

However, there is a silver lining, and one that I will mention again in Part III of this dissertation, when I outline a process for using technology to decrease bias in the law. As mentioned, we observed that fears of being accused of bias were correlated with decreased motivation to help. Furthermore, the less participants believed that bias actually impacted outcomes, the less they were willing to help. It is easy to design a system in which accusations of the sort that demotivate are not put forth, and it relatively easy to design a professional education

system in which attorneys are shown the facts regarding bias, are shown the extent to which it impacts their own legal decision making. Whether this might work to meaningfully decrease bias is something that can be glimpsed later in this dissertation.

Now, with this understanding of criminal defense attorneys, I turn in Chapter 3 to the other important player in the post-arrest and pretrial period: prosecutors.

Note

The research presented in this chapter was conducted by Joel Cooper and the dissertation author.

Chapter 3

Bias in Prosecutorial Decision Making

Racial disparities in U.S. criminal justice are especially acute in the context of drug offenses (Alexander, 2010; Human Rights Watch, 2000; Quigley, 2012). In 2016, there were more than 81,000 individuals in federal prisons for drug offenses, and 37.9% of them were Black, while 21.6% were White (Carson, 2018). Moreover, a number of reports have shown that the percentage of Black individuals imprisoned for drug crimes is not proportional to arrest rates (Chin, 2013). In other words, once arrested for a drug crime, Black individuals are more likely to go to prison. These disparities in drug convictions exist in spite of the fact that rates of using and selling drugs are similar for White and Black individuals (Miech, Johnston, O'Malley, Bachman, & Schulenberg, 2014; Substance Abuse and Mental Health Services Administration, 2017).

The stakes of improper or unjust convictions are high. The American Bar Association identified more than 38,000 punitive provisions that apply to people convicted of crimes, including loss of public housing, welfare assistance, and occupational licenses (Blumstein & Nakamura, 2012), as well as loss of financial aid for higher education and even social security benefits (Jacobs, 2015). Felons may lose the right to vote (Jacobs, 2015). Incarceration itself poses significant health risks (Haney, 2003; Ludwig, Cohen, Parsons, & Venters, 2012). Inmates routinely face sexual abuse (Beck & Harrison, 2007), and they often fail to receive proper mental health care and treatment (Oser, 2009; Rich, Wakeman, & Dickman, 2011). After release, something as essential and as seemingly quotidian as finding employment becomes difficult if not impossible (Young & Petersilia, 2016).

Given the salience of unjust arrests and unjust trials (Johnson, Blume, Eisenberg, Hans, & Wells, 2012; Liebman & Clarke, 2011; Liptak, 2018; Pierce & Radelet, 2011; Southall, 2017),

these stages of the criminal justice process receive significant attention from scholars and policy experts (Levinson, Bennett, & Hioki, 2017). But racial bias that emerges between the two poles of arrest and trial is undoubtedly of utmost important. Approximately 95-97% of convictions are obtained by guilty plea (Cohen & Kyckelhahn, 2010; Durose & Langan, 2007; Pastore & Maguire, 2003).[10] In short, almost all criminal cases resolve in the interstice between arrest and trial. The rules governing this period are unclear, and the primary authorities are largely immune from oversight. If one seeks to lessen racial disparities in incarceration rates, this is the period that needs to be better understood and better addressed.

Prosecutorial Power

Prosecutors are arguably the most influential actors in the U.S. criminal justice system (Gilbert & Johnson, 1996; Kutateladze, Lynn, & Liang, 2012; Miller, 2004; Stith, 2008). They have the power to drop a case, to oppose bail or to recommend a certain level of bail, to add or remove charges and counts, to offer and negotiate plea bargains, to propose deals in which the defendant offers testimony against another defendant in exchange for leniency, and to recommend sentences.

Prosecutors also operate with nearly unchecked discretion and are able to execute the above powers in the absence of meaningful oversight (Bibas, 2009; Davis, 2009, 2017; Smith & Levinson, 2012; Wright, 2017). Even courts have recognized that charging discretion is "almost limitless" (State v. Kenyon, 270 N.W.2d 160, 164 (Wis. 1978)). This discretion stems from the recognition, which most states make, that, pursuant to the federal constitutional model, the Executive Branch, out of which prosecutors operate, has "exclusive authority and absolute

[10] Plea bargaining is a process wherein a defendant receives less than the maximum charge in exchange for an admission of guilt or something functionally equivalent to guilt prior to the case going to trial (Cohen & Reaves, 2006).

discretion to decide whether to prosecute a case" (Greenlaw v. United States, 554 U.S. 237, 246 (2008) (quoting Nixon, 418 U.S. at 693)). In other words, when a new case arrives on a prosecutor's desk, he or she has the power to be harsher on certain defendants—and more lenient on others.

Prosecutorial Decision Making

In spite of the importance of prosecutorial decision making, empirical research analyzing racial bias in prosecutorial decision making has been relatively scarce (Kutateladze, Lynn, & Liang, 2012). Moreover, the results that have been obtained are somewhat inconclusive. A 2012 meta-analysis that reviewed 34 studies found that, in most but not all instances, race affected case outcomes, even after controlling for a host of non-race legal and extra-legal factors (Kutateladze, Lynn, & Liang, 2012). The results were not always beneficial to White individuals and detrimental to minorities, however: it seemed to depend, in some studies, on the specific offense or the specific point in the criminal justice process that was being evaluation (Kutateladze, Lynn, & Liang, 2012). For example, Spohn and Spears (1996) found that while prosecutors were more likely dismiss Black defendant and White victim sexual assault cases, once such a case is commenced, prosecutors are more likely to ask for harsher penalties.

On one side of the debate, there is accumulating evidence of racial bias in prosecutorial decision making. A 2012 study of cases in Washington State found that prosecutors are less likely to charge White suspects than Black suspects (Research Working Group of the Task Force on Race and the Criminal Justice System, 2012). Similarly, Radelet and Pierce (1985) reviewed prosecutorial decision making in over 1,000 Florida homicide cases and found a combination of harsher treatment of Black defendants and more lenient treatment of White defendants. Analysis of criminal cases disposed of in 2010-2011 in New York City found that Black and Hispanic

defendants were less likely than White defendants to have their charges decreased by the prosecutor (Kutateladze, 2017). Berdejô reviewed almost 50,000 plea bargains in Wisconsin, finding that White defendants were 25 percent more likely than Black defendants to have their charges reduced or dropped altogether (Berdejô, 2018). Prosecutors have been shown to be less likely to offer Blacks a plea bargain, less likely to reduce charge offers for Black defendants, and more likely to offer Black defendants plea bargains that include prison time (Albonetti, 1990; Kutateladze, Andiloro, & Johnson, 2016; Kutateladze, Andiloro, Johnson, & Spohn, 2014; McDonald & Raphael, 2017). Defendants who were Black, young, and male fare especially poorly (Albonetti, 1990; Kellough & Wortley, 2002; Kutateladze, Andiloro, & Johnson, 2016). In a comprehensive study of federal cases that resolved in 2006-2008, Rehavi and Starr (2014) found that Black defendants received sentences that were 10 percent longer than those of similarly-situated Whites and that this disparity was explained by prosecutors' initial charging decisions.

In addition, it is worth noting that explicit racial stereotyping has long been a feature of prosecutorial trial behavior. Striking Black jurors used to be based on explicit racism. For instance, in the 1960s and 1970s, the Dallas County District Attorney's Office used a trial manual that stated, "Do not take Jews, Negroes, Dagos, Mexicans or a member of any minority race on a jury, no matter how rich or how well educated" (*Miller-El v. Cockrell*, 2003). While the Court ruled in *Batson v. Kentucky* (1986) that prosecutors cannot strike a potential juror based solely on race, it is widely acknowledged that prosecutors continue to sidestep the *Batson* Court's admonition (Page, 2005; Sommers & Norton, 2007). Supreme Court Justice Stephen Breyer wrote that he was not surprised to learn that, despite *Batson*, discriminatory use of peremptory challenges remains common (*Miller-El v. Dretke*, 2005).

On the other side of the evidentiary debate, there are studies that have failed to find racial bias in prosecutorial decision making. Franklin (2010) used a nationally representative sample of felony drug cases and found no effect of race on prosecutors' decisions to dismiss charges. Albonetti and Hepburn (1996) found no evidence that Black individuals were less likely than White or Hispanic individuals to be diverted into a treatment program in lieu of criminal prosecution. When it comes to charge reductions, two studies analyzing cases from the 1980s and 1990s found no effect of race (Albonetti, 1992; Spohn & Horney, 1993). Shermer and Johnson (2010) reviewed federal cases that resolved in 2001 and found that offender race exerted no direct influences on federal charge reductions.[11]

This inconsistency in prior findings partly reflects that these studies looked at different jurisdictions and different time periods. Additional inconsistency can be explained by the fact that not all of these studies analyzed the precise same crimes. That is, drug offenders who are Black might be treated worse than drug offenders who are White, but perhaps burglars who are Black fare better than burglars who are White. In some sense, such nuances are unsurprising, as not all crimes are the same and stereotypes seem to vary across crimes, which is something I explore in Chapters 4 and 5. At present, however, I focus on drug crimes, which are highly and consistently stereotyped.

Current Research

Given the increasing attention paid to racial bias, a number of offices have commenced interventions in an attempt to lessen racial disparity (see McKinley, 2014). Whether such

[11] However, in spite of the careful work represented by these studies, the fact that Black and White individuals are "treated the same" may be camouflaged by the fact that the initial sets are different. For instance, if the threshold for charging Black individuals is lower, then even if an office dismisses 40% of both Black and White individuals, it still will have dismissed a greater proportion of White (vs. Black) defendants accused of relatively more severe crimes.

interventions are working is not yet known (Baker, 2018), but the mere presence of such interventions increases the need for evaluations of recent case outcomes. Results from a decade ago might not say much about what needs to be done today. The present study focuses exclusively on cases that resolved in a large U.S. city in 2016.

The present research population holds two additional strengths. First, the population is limited to those who were charged with felony drug crimes. As discussed above, drug incarceration is an area of extreme disparity, and research into prosecutorial decision making in this same area stands to account for much of the statistical disparity observed in the justice system as a whole. Second, much of the existing literature on racial disparity in prosecutorial decision making has focused on discrete moments in the criminal justice process, which may fail to accurately represent the process (Rehavi & Starr, 2014). The present study is more expansive, as we focus on charge reductions, a dynamic measure that takes into account what happened to a case in terms of both the initial arrest and the final disposition reached by plea bargain.

Method

Population

A prosecutor's office in a large, urban locality provided our research team with case information for 1,711 felony drug cases that resolved during 2016. By "felony," we mean that the prosecutor's initial charging decision included at least one felony charge. Eighty-nine percent of the defendants were male, and 11% were female. Average age was 38 ± 13 years old. One hundred and ninety defendants were White (11%), 823 were Black (48%), 662 were Hispanic/Latinx (39%), 25 were Asian (1%), and 11 were listed as "other" (1%).

Case Data

The data set contained a number of data points for each defendant's case. These data points included characteristics specific to the defendant, including age, gender, race, education, number of dependents, whether or not the individual was homeless, employed, or in school, and the defendant's residential zip code. The data also included characteristics specific to the arresting officer and the prosecuting attorney. It included a detailed accounting of the defendant's criminal history, including conviction density.[12] It included information specific to the pending charge(s), such as the specific drug possessed and whether drug paraphernalia was found.

The primary dependent measure was the change in case outcomes from arrest to final disposition. Change was operationally defined as a charge reduction: from felony at arrest to misdemeanor/violation/infraction/dismissal at plea. Given that only 17 of the 1,711 cases resolved by trial, we excised those cases from our analyses.

Results

Preliminary Analyses

Preliminary analyses considered raw rates of case outcomes and charge reductions. At final disposition, 62% of White defendants received a felony and/or misdemeanor conviction (meaning that 38% of White defendants received either a violation, infraction, or a dismissal), while 76% of Black and 74% of Hispanic defendants received a felony and/or misdemeanor. Seventy-three percent of White defendants received a charge reduction (from arrest to final disposition), while 57% of Black and 53% of Hispanic defendants received the same. Logit regression yielded a significant negative effect of both Black ($p < .001$) and Hispanic ($p < .001$)

[12] See Tollenaar (2017) and Copas and Marshall (1998): this variable is computed as √{#convictions/(careerlength+1))}.

race on charge reductions. Analysis of variance across the three race categories revealed a significant difference, df = (2, 1672), $p < .001$. The Wald Test produced a chi-squared test statistic of 53.4, with three degrees of freedom and an associated p-value < .001.

Analysis of Similarly-Situated Defendants

For this analysis, we limited the population to the three largest racial groups: White, Black, and Hispanic (N = 1675), and we further limited the population to only those defendants who had no prior convictions (N = 501). We determined the rates of charge reductions across the three racial groups. Eighty percent (standard error: ± 4.2%) of White defendants received a charge reduction, while only 66% (± 3.5%) of Black and 61% (± 3.2%) of Hispanic defendants received a charge reduction. Logit regression yielded a significant negative effect of both Black ($p = .02$) and Hispanic ($p = .001$) race on charge reductions. Analysis of variance across the three race categories revealed a significant difference, df = (2, 499), $p = .003$. The Wald Test produced a chi-squared test statistic of 56.5, with three degrees of freedom and an associated p-value < .001.

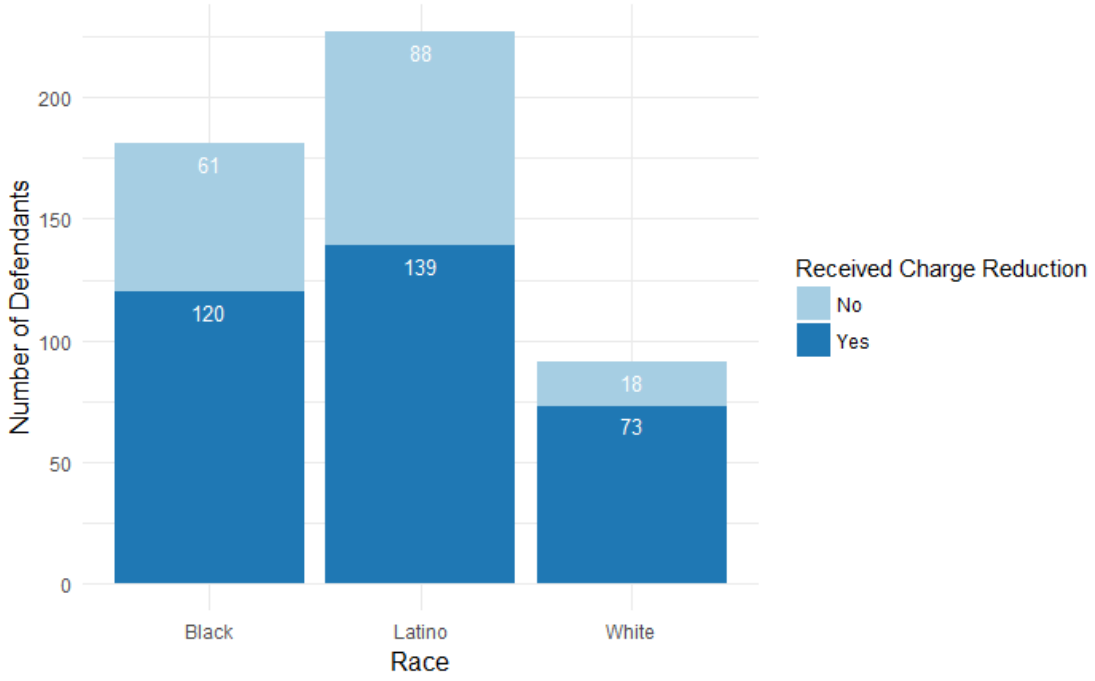


Figure 3.1 Charge reductions for defendants with no prior convictions, by race. Charge reduction is defined as when a defendant was cited with a felony at arrest but ultimately pleaded guilty to a misdemeanor or less.

We also analyzed charge reductions for the largest bin of similarly-situated defendants, that is, those who had the same top arrest charge seriousness (Felony B), same gender (male), same education level (high school), were not students, did not have dependents, and had the same criminal history (between 0.4 and 0.8 on the conviction density scale). For these 259 defendants, 65% (± 11.6%) of White, 51% (± 4.2%) of Black, and 36% (± 4.8%) of Hispanic defendants received a charge reduction. Logit regression revealed a significant negative effect of the defendant being Hispanic ($p = .01$), but the effect of the defendant being Black did not rise to significance ($p = .10$), although the effect was negative. This sample suffered from a small population of White defendants (N = 17). Analysis of variance across the three race categories revealed a significant difference, df = (2, 257), $p = .03$. The Wald Test produced a chi-squared test statistic of 43.4, with three degrees of freedom and an associated p-value < .001.

Analysis of the Most Significant Case Features

We fit a generalized linear model to the arrest data, including the features related to the defendant and/or the alleged crime: age, gender, race, number of dependents, education level, student status, employment status, housing status (homelessness), criminal history, and the primary drug for which the defendant was arrested. We used a binary binomial distribution for Y in the binary logistic regression, where Y, our outcome measure, was whether or not the defendant received a charge reduction. Stepwise model selection by AIC yielded a model in which only four of these features showed significant effects (Table 3.1 and Figure 3.2). Null deviance for the model was 2207.8 on 1618 degrees of freedom, and residual deviance was 2129.1 on 1602 degrees of freedom. AIC was 2163.1. In the figure that follows, the coefficients of logistic regression reflect the odds ratio of the probability.

Table 3.1 Significant features from logistic regression of arrest data with charge reduction as the binary outcome measure. A negative estimate means that the defendant is less likely to receive a charge reduction, while a positive estimate means that the defendant is more likely to receive a charge reduction.

Feature	**Estimate**	**Std. Error**	**z-value**	***p*-value**
White race	.51	.20	2.58	.01
Male sex	-.34	.17	-2.06	.04
Dependents	-.14	.06	-2.50	.01
Primary Drug = heroin	-1.03	.23	-4.43	<.001
Primary drug = powder cocaine	-.77	.26	-3.02	.003
Primary drug = crack cocaine	-.76	.24	-3.20	.001

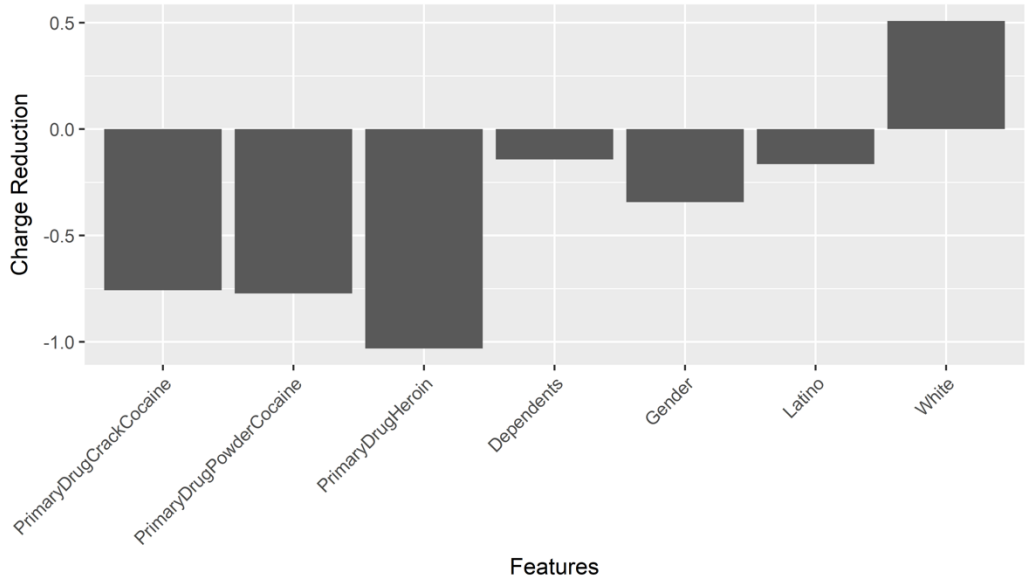


Figure 3.2 Significant features (and also Latino race, which had a *p*-value of .16) from logistic regression of arrest data with charge reduction as the binary outcome measure. A negative estimate means that the defendant is less likely to receive a charge reduction, while a positive estimate means that the defendant is more likely to receive a charge reduction. The baseline of 0 reflects Black race.

Discussion

The findings consistently showed racial disparate rates in case outcomes, with White defendants (vs. Black and Hispanic defendants) benefiting. For the preliminary analyses, at final disposition, 62% of White defendants received a felony and/or misdemeanor conviction, while 76% of Black and 74% of Hispanic defendants received a felony and/or misdemeanor. Seventy-three percent of White defendants received a charge reduction (from arrest to final disposition), while 57% of Black and 53% of Hispanic defendants received the same. Logit regression yielded a significant negative effect of both Black and Hispanic race on charge reductions. For the defendants with clean criminal records, 83% of Whites received a charge reduction, while only 66% of Blacks and 61% of Hispanic defendants received a charge reduction. Logit regression yielded a significant negative effect of both Black and Hispanic race on such charge reductions.

For the similarly situated defendants, 65% of Whites, 51% of Blacks, and 36% of Hispanic defendants received a charge reduction. Finally, for the analysis of the most significant case features, stepwise model selection by AIC yielded a model in which only four of the features showed significant effects, with race being one of them. In particular, the results revealed that, controlling for all relevant features, defendants who were White were significantly more likely to receive a charge reduction.

What does this mean? As is evident in Table 3.1, primary drug and defendant criminal history were not significant. What mattered were specific characteristics of the defendant. In short, the crime itself and the defendant's criminal history mattered less than the demographic characteristics of the defendant, especially race.

Understanding Prosecutorial Bias

As tempting as it is to think that each case is unique and that prosecutorial work is a nuanced undertaking, prosecutors often operate with limited information about what actually transpired, and this may lead to reliance on extralegal influences, such as racial stereotypes (Kutateladze, Lawson, & Andiloro, 2015). An analysis of 2010 and 2011 felony drug cases in New York City found that, of numerous evidentiary factors considered, almost none mattered for predicting sentence offers (Kutateladze, Lawson, & Andiloro, 2015). In addition to limited information, prosecutors also operate under time constraints, another factor that has been shown to lead to reliance on extralegal factors (Albonetti, 1991; see also Kahneman, 2011, for a discussion of reliance on heuristics). Given time and information constraints, legal decision makers are unable to accurately assess an alleged offender's dangerousness and culpability, and this makes them liable to overreliance on stereotypes, such as those that attach to African Americans (Hawkins, 1981; Steffensmeier, Ulmer, & Kramer, 1998).

Despite such informational and time constraints, prosecutors are instructed to make nuanced determinations. For instance, the Department of Justice instructs federal prosecutors, in deciding whether or not to pursue resolution through plea bargaining, to consider a defendant's remorse or contrition and also the interests of the victim, including any effect upon the victim's right to restitution (U.S. Department of Justice, 1991). Both of these factors are ones likely to be impacted by pervasive racial stereotypes. Perceptions of remorse and contrition are highly subjective, and we know that race may impact related perceptions. For example, across seven studies, Wilson, Hugenberg, and Rule (2017) found that people perceive Black men as both bigger and more physically threatening than White men, and these misperceptions occurred as the result of both bottom-up cues of racial prototypicality and top-down information. The authors further found that these misperceptions were used to justify beliefs that Black men should be subjected to more aggressive law enforcement measures. Goff, Eberhardt, Williams, and Jackson (2008) reviewed newspaper reports for 600 cases that took place from 1979-1999 in Philadelphia, finding that, when the defendant was Black, there was a fourfold increase in the number of dehumanizing (animalistic) references—and those references strongly correlated with the defendant receiving the most severe punishment available. Whether such beliefs as these affect prosecutors and impact their determinations of remorse or contrition (or even likelihood of recidivism, which is tied up in the same determination), is an empirical question, as is the question of whether defendants are treated more harshly when the defendant/victim dyad is Black/White (Bowers, Sandys, & Brewer, 2004; see also Jennings, Richards, Smith, Bjerregaard, & Fogel, 2014). If prosecutors, on account of such biases, tend to conclude that Black males are more intimidating, more aggressive, and perhaps less remorseful for their criminal offenses, and if prosecutors are likely to be harsher on Black defendants when the victim is White, then the

very guidelines in place for prosecutorial decision making will facilitate racially biased outcomes.

Limitation

This study is based on a single jurisdiction and may not be generalizable to other jurisdictions. That said, the cases reviewed were from a large, urban center that is not in the U.S. south and is not generally considered an area beset by racial bias or animus.

Conclusion

The findings consistently showed disparate rates of charge reductions for one of the most common and pervasive offense types—drug crimes—with White defendants (vs. Black and Hispanic defendants) benefiting. Even after controlling for a nearly exhaustive list of factors, race still was driving these case outcomes, suggesting significant bias by prosecutors.

Thus, in Part I of this dissertation, I have added to the literature on bias in criminal defense, finding significant disparities in both criminal defense attorneys and prosecutors. Given their centrality to the criminal justice process, we might begin to wonder two things. First, what could be driving such bias on a fundamental, basic level? Second, what can be done about it? In Parts II and III of this dissertation, I take up these questions.

Part II

What Drives Bias in Legal Outcomes?

Part I of this dissertation provided insight into bias by legal actors, specifically criminal defense attorneys and prosecutors. Part II expands the scope, focusing on criminal stereotypes in the broader U.S. population and exploring the complexities of these stereotypes and their effects.

To date, researchers have provided significant insight into criminal stereotypes, generally concluding that, in the United States, criminality is associated with Black skin color. For example, when Eberhardt and colleagues asked police officers the question, "Who looks criminal?", the officers chose more Black faces than White faces (Eberhardt, Goff, Purdie, & Davies, 2004). The general conclusion that criminality is associated with Blackness and Blackness is associated with criminality has become a truism in social psychology and related fields (see Allport & Postman, 1947; Correll, Park, Judd, & Wittenbrink, 2002; Eberhardt, 2019; Payne, Lambert, & Jacoby, 2002).

But, in thinking about criminality in general, have we overlooked nuances that attach to subtypes of criminality? There are many different crimes that one can commit, and the stereotypes for these crimes do not all reflect a general Black-crime association. Across the four significant scholarly forays into stereotypes of criminal subtypes, divergence from the Black-crime association has been found, with certain crimes associated with White individuals—and a few other crimes associated with other ethnicities, such as Middle Eastern/Arab individuals (Esqueda, 1997; Gordon, Michels, & Nelson, 1996; Skorinko & Spellman, 2013; Sunnafrank & Fontes, 1983). However, those working on criminal subtypes have generally not offered a theoretical account as to what is driving such subtype associations. There has been a dearth of exploration of whether there exists—beyond or beneath the impact of base rates and/or social

learning—a more fundamental understanding or misunderstanding that precipitates these divergences in stereotypes. In Part II of this dissertation, I suggest that, indeed, the standard psychological story has overlooked an important wrinkle—that is, moral stereotyping—that informs stereotypes of criminal subtypes and impacts criminal justice decision making. In particular, what is driving criminal subtype stereotypes and their effects might be more insidious and deeply ingrained than previously thought.

Chapter 4

Identifying Stereotypes of Criminal Subtypes via Reverse Correlation

In this chapter, I further explore bias in legal thought, but I do so from a more fundamental level. Specifically, I suggest that the standard psychological story has overlooked an important wrinkle—that is, moral stereotyping—that informs stereotypes of criminal subtypes and impacts criminal justice decision making. By the end of this chapter, we should be able to better understand how moral stereotypes operate in the background, driving much of what we have heretofore considered as solely criminal stereotypes.

Introduction

Research during the past several decades has identified a persistent relationship between race and criminality (Avery & Cooper, 2020). In particular, Black individuals are more often identified with crime, and crimes are more likely to be associated with Black perpetrators (Allport & Postman, 1947; Correll, Park, Judd, & Wittenbrink, 2002; Payne, Lambert, & Jacoby, 2002). But this stereotype stands in contrast to base rates of criminal offending, as 27% of arrestees in the United States are Black (FBI, 2019). Moreover, there is general scientific agreement that the stereotypes for auto thief and burglar are Black, even though incarceration demographics show that prisons are populated with more White (vs. Black and also vs. Hispanic) auto thieves and burglars (Bronson and Carson, 2019). In addition, for most crimes, offense rates are very nearly split at 1/3 each for White, Black, and Hispanic defendants (for 2017 incarceration rates, see Bronson & Carson, 2019; for 2018 arrest rates, see Federal Bureau of Investigations, 2019). In addition, in forming stereotypes, it is unclear if one would look to incarceration demographics or overall rates by group, and these two may very well conflict with each other. Even more confounding is the fact that, regardless of the sample on which one

chooses to focus, any resultant stereotypes could be generally misleading, since the differences in offense rates across groups and the chances of one being victimized are so low as to make it exceedingly difficult to predict what any individual will do.

But the stereotype does align with a wealth of social psychological research suggesting that Black individuals are often erroneously thought to be aggressive, violent, and criminal (Devine, 1989; Devine & Elliot, 1995; Sagar & Schofield, 1980). As Eberhardt and colleagues argued, what is at work is a prevalent societal stereotype that casts Black individuals as the "prototypical embodiment" of crime, regardless of whether there is personal endorsement of the stereotype (Eberhardt et al., 2004).

Going beyond race, sex also is a demographic feature that undeniably impinges upon crime stereotypes. It is widely believed that crime is a male endeavor (Thiem, Neel, Simpson, & Todd, 2019). Indeed, the base rates overwhelmingly support such a belief. For all crimes combined, in 2018, 73% of those arrested were male. Moreover, in 2017, 93% of state and federal prisoners were male (Bronson & Carson, 2019), with the only notable exception being prostitution, which was 64% female and which made up a mere .3% of all arrests in 2018 (Federal Bureau of Investigations, 2019).

Criminal Subtypes

Research on criminal subtypes has added nuance to the association of Blackness with crime. A number of studies have shown that Blacks and Whites are differentially associated with certain types of crimes. Esqueda (1997) found that armed robbery, mugging, murder and shoplifting were thought to be committed by Black individuals, whereas embezzlement, fraud, counterfeiting and bribery were identified with White individuals. Skorinko and Spellman (2013)

found that Black Americans were uniquely associated with such crimes as assault, auto theft, rape, burglary, and murder, while European Americans were uniquely associated with crimes that include arson, fraud, embezzlement, and drunk driving. There was no single explanation about how these groupings formed, although some suggested that there was a Black/White parsing that mirrored the blue collar/white collar crime distinction (Gordon, Michels, & Nelson, 1996; Sunnafrank & Fontes, 1983). Skorinko and Spellman (2013) found that, relative to Whites, Blacks were associated with crimes of violence, but this distinction did not fully explain the stereotypical differences. In this chapter, without rejecting the potential relevance of these distinctions, we explore an underlying factor that we believe plays an important part in distinguishing criminal subtypes, namely, the moral dimension.

Morality and the Black Criminal Stereotype

We believe that at least one of the dimensions that precipitates the distinction between stereotypically White and stereotypically Black crimes is the moral dimension. Specifically, we believe that crimes viewed as fundamentally immoral are more likely to be associated with Black individuals, while crimes viewed as less immoral are more likely to be associated with White individuals. We are guided in this notion by the heuristic introduced by Alter, Stern, Granot, and Balcetis (2016). Across six studies, Alter et al. (2016) found that people used a "bad is Black" heuristic, such that they assumed immoral acts were committed by people with darker skin tones. This was true in newspaper articles, it was true in examination of surveillance footage, and it was true when attributing good/bad actions to different people. They also found that the higher the degree of imputed immorality, the darker was the associated shade of skin tone.

Alter and colleagues' research accords with previous work finding that people are quicker to associate Blackness with badness and Whiteness with goodness, and they struggle to form the opposite associations i.e., Blackness with goodness and Whiteness with badness (Meier, Robinson, & Clore, 2004; Sherman & Clore, 2009). The bad is Black heuristic has been argued for in some form or another for years, especially in the context of drug use. For example, at the onset of the current opioid epidemic, there is evidence that minority substance users were considered irresponsible and criminal while White substance users were viewed as suffering from mental disorders and in need of treatment (Netherland & Hansen, 2016).

Within criminal law, a moral classification of crime is already well established (Black, 1990), and it accords with the moral dimension that we have been discussing. *Malum prohibita* crimes are those that, while prohibited by statute, are not perceived to be inherently wrong. As Greenfield (1921) summarized: they are "offenses not involving moral turpitude" (p. 493). *Malum in se*, by contrast, refers to crimes that are inherently wrong, where wrong might be thought of in reference to principles of natural law (Black, 1990).

Most crime types may readily be distributed into one of the two categories (see *District of Columbia v. Colts*, 282 U.S. 63, 73, 1930). Moreover, the distinction is useful in that it roughly corresponds to the social scientific understanding of morality and immorality. There is controversy about the factors that underlie people's judgments of moral or immoral (Haidt, 2001; Schein, Ritter & Gray, 2016), but people's judgments about the kinds of acts they find immoral are relatively consistent (Schein and Gray, 2015). We propose that this moral distinction serves to distinguish crimes perceived as stereotypically Black from those perceived as stereotypically White.

The Plan of this Research

In the present research, we use a novel method for assessing the relationship between *mala in se/prohibita* crimes and race. We hypothesize that participants will show systematic associations between race and crime types, and that those associations will reflect a moral dimension. That is, crimes considered *mala in se* will be associated with Black offenders, and crimes considered *mala prohibita* will be associated with White offenders.

Research on criminal subtypes has largely relied on questionnaire methodology (e.g., see Esqueda, 1997; Skorinko & Spellman, 2013), which raises the longstanding problems of honesty and social desirability when asking about sensitive matters such as racial stereotypes (Dovidio & Gaertner, 2000; Krumpal, 2013). In the current work, we make use of recent developments in social scientific techniques to visualize stereotypes of criminal subtypes. The data-driven method that we use—reverse correlation to produce classification images (Dotsch & Todorov, 2012)—overcomes many of the difficulties that have characterized past studies using explicit measures.

Reverse correlation takes judgments of randomly varied stimuli and produces proxies of mental representations (Brinkman, Todorov, & Dotsch, 2017). More specifically, participants are shown two different faces, selected at random from a large data base of images, and are asked to select the face that best resembles a given prompt, e.g., a bank robber. This is repeated many times, and the responses of all participants are combined to produce an average mental representation of the bank robber. This procedure has been validated within a number of studies, including ones focused on social face perception (Dotsch & Todorov, 2012), stereotype content judgments (Imhoff, Woelki, Hanke, & Dotsch, 2013), and dehumanization (Kunst, Kteily, & Thomsen, 2019).

After generating the stereotypes in Study 4.1, we ran a follow-up study in which a separate sample of participants are asked to rate and categorize the resulting images. We then used a clustering machine learning algorithm to identify patterns in the results and render data-driven support for our prediction that the moral/immoral crime distinction would mirror a Black/White defendant divide.

Study 4.1: Visualizing the Subtypes

Participants

Three hundred and fifty-seven undergraduate students participated in Study 4.1, As we sought to capture the American stereotype of these subtypes, we did not include any international students in the population. Thirty-four percent were male, 66% were female, and less than 1% selected transgender, other, or preferred not to disclose. Mean age (± SD) of the participants was 20 ± 1.4 years. The racial breakdown was 43% White/Caucasian, 10% Black/African American, 8% Hispanic/Latinx, 32% Asian/Pacific Islander, 2% Middle Eastern/Arab, and 5% selected "other."

Methods

All of the experiments that we present in this chapter were approved by Princeton University's Institutional Review Board. In this first study, we used a noise-based reverse correlation technique to create classification images, a process that proceeds by presenting two blurry images and having participants choose which of the two images most resembles a target condition (e.g., which of these two images looks most like a murderer?). This process is repeated many times, the images chosen by participants are compiled, and these are combined to produce an overall outcome image.

Consistent with Dotsch et al. (2008), we gathered face images by combining composites of images of adults of different races. As is typical for this procedure, we averaged composite face images of male and female adults who racially self-identified as being from different races, including African, East Asian, West Asian, Caucasian, and multi-racial (n = 16 in total; mean age = 26.23 years, range age = [19,47]) (see DeBruine, 2016). We conducted this process using PsychoMorph (see Tiddeman, Burt, & Perrett, 2001). The resultant stimulus represented the average of all the composite face images and was ambiguous in sex and race (see "Baseline Image" in Figure 4.2). We then grayscaled, cropped, and blurred this averaged image using a Gaussian blur with 5-pixel radius (see Dotsch, Wigboldus, Langner, & van Knippenberg, 2008). This stimulus image was then superimposed with random noise with two opposite greyscale values in each corresponding pixel, resulting in distinct images. Each noise overlay was the average of multiple 2D sinusoid patches with various cycles per patch. This process was repeated a total of 500 times, yielding 500 noise-superimposed face pairs.

These pairs of faces made up the trials that participants completed. For each trial, we presented participants with two faces and asked them to select the one that looked most like the given condition (a burglar, a murderer, and so on). For example, a participant in the burglar condition would see a pair of faces and would select the one that looked most like a burglar, and the participant would repeat this same task a total of 500 times. This process took approximately 25 minutes. After all participants had completed their 500 selections, a resultant image, called the classification image (CI), was computed by averaging the images selected. The CI represents an approximation of the average mental representation that participants had of the category.

Participants were asked to make their judgments on one of nine categories. These included categories in which the specific crime was unspecified but the infraction was either a

malum prohibitum crime (recall that this a crime that is wrong because it is prohibited) or *malum in se* crime (wrong in and of itself). Specifically, for *malum prohibitum*, participants were told, "*Malum prohibitum* means a crime that is wrong simply because it is prohibited by a statute, not because it is inherently wrong." For *malum in se*, they were told, "*Malum in se* means a crime that is wrong because it is inherently wrong."

Also included as categories were law-abiding person, criminal, insider trader, burglar, structuring offender, assault offender, and murderer. We chose these crimes because burglary, assault, and murder seem like *mala in se* offenses, while insider trading and structuring (structuring is the act of breaking up financial transactions to get around federal reporting requirements) are two paradigmatic examples of *mala prohibita* offenses (see Kahan, 1997; Travers, 1995).

To verify that the specific crimes were indeed so perceived as *malum prohibitum* or *malum in se*, we conducted a pilot study with 100 U.S. citizens living in the U.S. recruited from Positly. The participants were shown the five specific crimes (burglary, assault, murder, insider trading, structuring) in random order, and for each crime, they were asked where it fell on a *malum prohibitum/malum in se* scale (with the terms defined the same as in the preceding paragraph), where -10 was "definitely *prohibitum*" and 10 was "definitely *in se*." They also were asked where it fell on an immoral scale, where 0 was "not at all immoral" and 10 was "completely immoral."

The results supported both the predicted *malum prohibitum/in se* distinction and the attendant immorality attributions. Insider trading and structuring were both identified as on the *malum prohibitum* side of the scale, with means of -1.45 and -2.00, respectively, while burglary,

murder, and assault scored 5.59, 8.05, and 6.21, respectively (see Figure 4.1). The mean immorality ratings for insider trading, structuring, burglary, murder, and assault were 6.53, 5.70, 8.42, 9.19, 8.64, respectively.

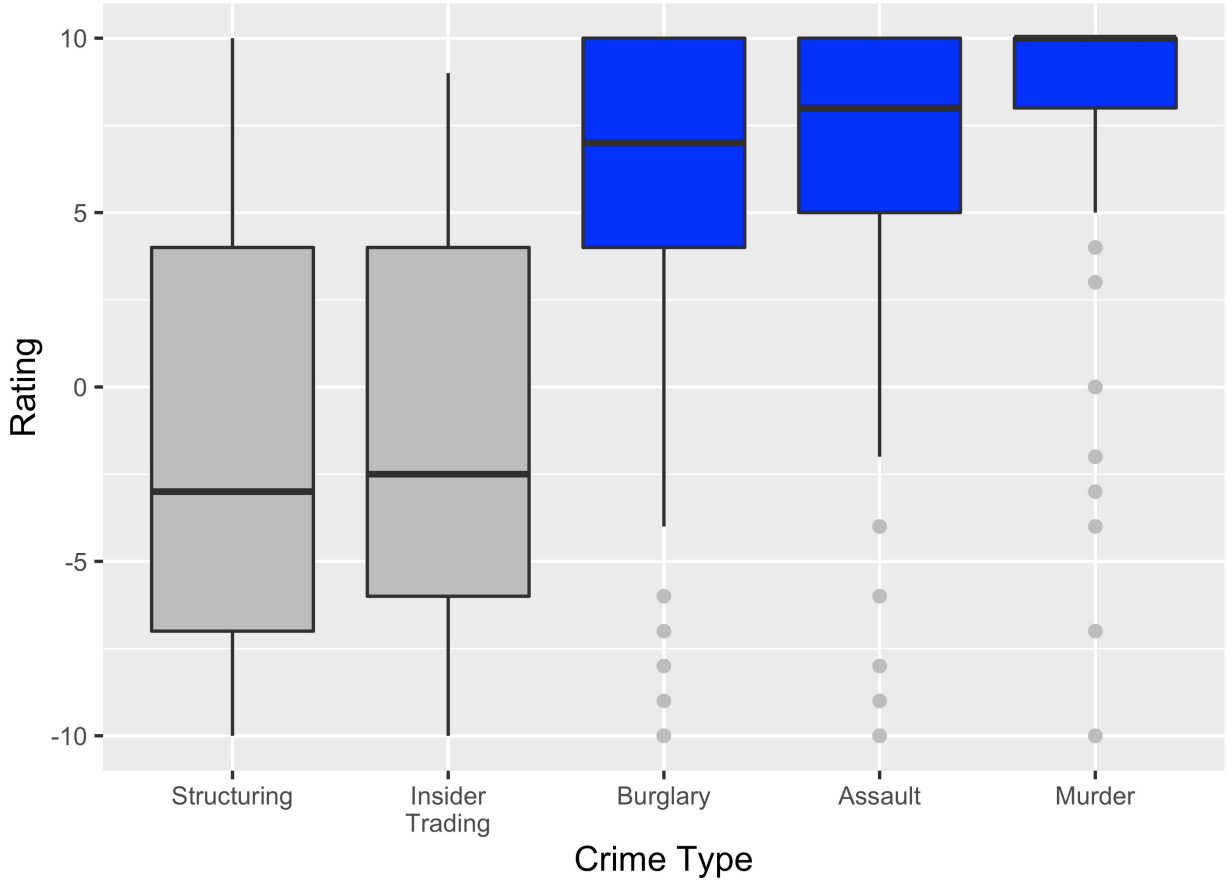


Figure 4.1 Ratings of the five specific crimes subsequently used in our reverse correlation procedure. 10 indicates definitely *malum in se* and -10 indicated definitely *malum prohibitum*.

Results

Figure 4.2 displays the results of the reverse correlation procedure. The baseline stimulus is at top; beneath the stimulus are the nine classification images, which represent the

visualizations of the stereotypes for the listed categories. In the succeeding study, a new group of participants will assess the race and sex of the face depicted for each crime.

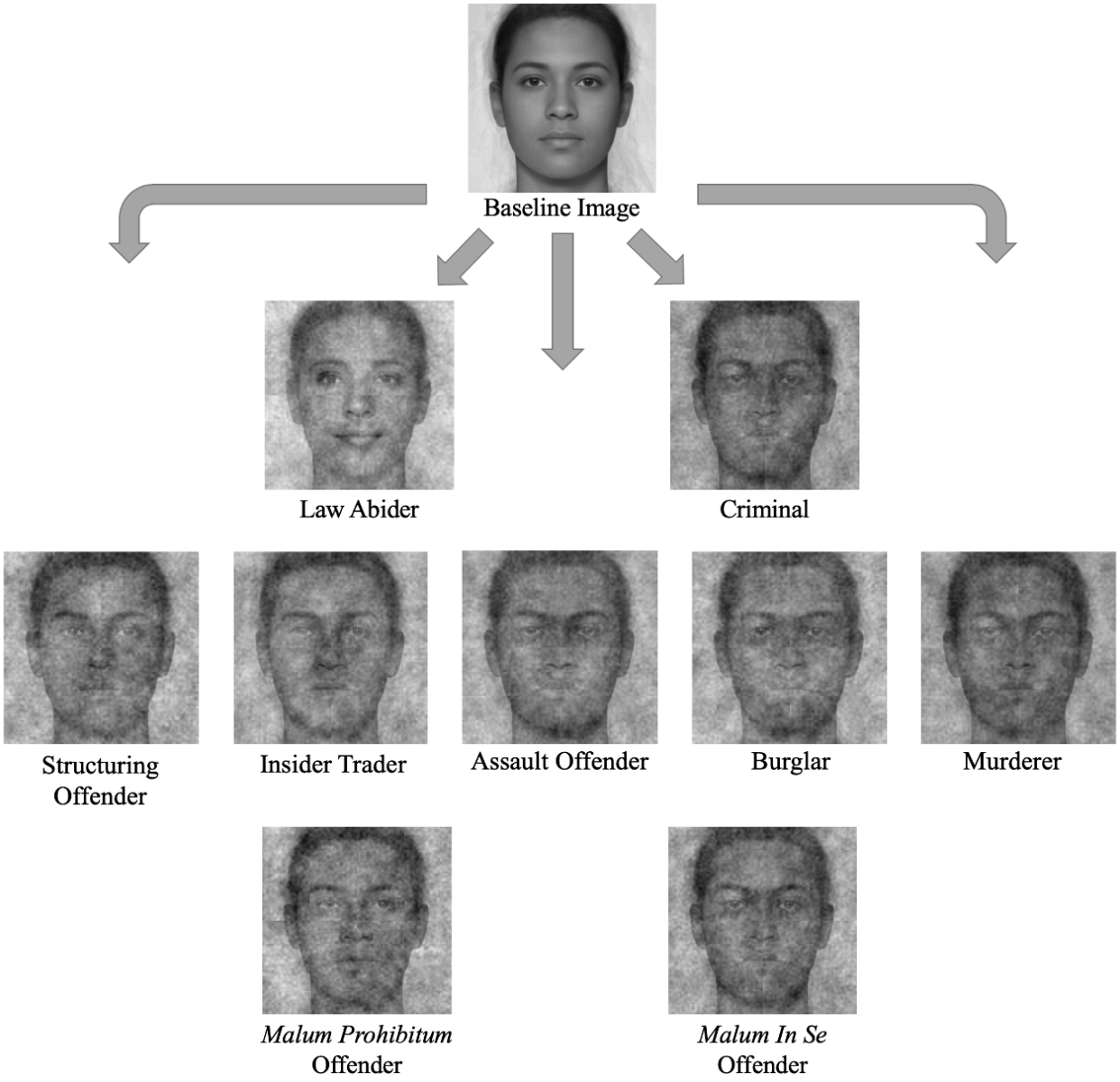


Figure 4.2 Classification images of stereotypes of criminal subtypes.

Summary of Study 4.1

In Study 4.1, across the pilot and the main experiment, we were able to predict the stereotypes of various *mala prohibita* and *mala in se* crimes. By mere visual inspection, these results appear to reflect our main hypothesis: that White males are assumed to commit *mala prohibita* crimes, while Black males commit *mala in se* crimes. Whether this appearance is scientifically true is an empirical question, one that we take up in Study 4.2.

Study 4.2: Rating and Grouping the Subtypes

In order to test our hypothesis that the "bad is Black" heuristic undergirds the formation of criminal subtype stereotypes, we needed to rate and group the classification images presented in Figure 4.2. In particular, we hypothesized that the initial parsing would be in line with the general association of crime with males. Thus, the only condition that should tend towards a female stereotype would be that of the law-abiding person. However, within criminality, criminal subtypes would parse along a racial (White/Black) line in accordance with the *malum prohibitum*/*malum in se* distinction. That is, participants would imagine a criminal who is not immoral as a White male and a criminal who is immoral as a Black male.

Methods

To examine this hypothesis, we tested the images produced along a few key dimensions, and we used k-means clustering, with the silhouette method of determining the number of groups (see Kaufman & Rousseeuw, 1990; for an example of how this technique is implemented in R, see https://uc-r.github.io/kmeans_clustering) to cluster the visualizations.

Specifically, participants viewed all nine classification images, presented in random order, and for each one they indicated where they believed the image fell on a sex scale and a race scale. The sex scale was a ten-point Likert, with "definitely male" on one end and

"definitely female" on the other end. The race scale was a ten-point Likert, with "definitely White" on one end and "definitely Black" on the other end.

Two hundred participants from Prolific were recruited, and 198 remained after exclusions. All were United States citizens who were living in the United States. Forty-five percent were male, 51% were female, 1% were transgender, 2% selected "other," and the rest preferred not to respond. Mean age (± SD) of the participants was 32 ± 11 years. The racial breakdown was 69% White/Caucasian, 7% Black/African American, 7% Hispanic/Latinx, 13% Asian/Pacific Islander, <1% Native American, <1% Middle Eastern/Arab, 3% selected "other," and the rest preferred not to answer.

Results

For the sex scale, all classification images were overwhelmingly considered male except for one: the law-abiding CI. (See Figure 4.3; see also the Appendix.) For the race scale, the law-abiding person was considered White and the general criminal (no crime type specified) was considered Black. For the five subtype stereotypes, the *malum prohibitum* offenders (structuring offender and insider trader) were considered White, and the *malum in se* offenders (burglar, assault offender, murderer) were considered Black. Finally, for the *malum prohibitum* and *malum in se* categories themselves, the *malum prohibitum* offender was considered, on average, White, while the *malum in se* offender was considered Black. (See Figure 4.4; see also the Appendix.)

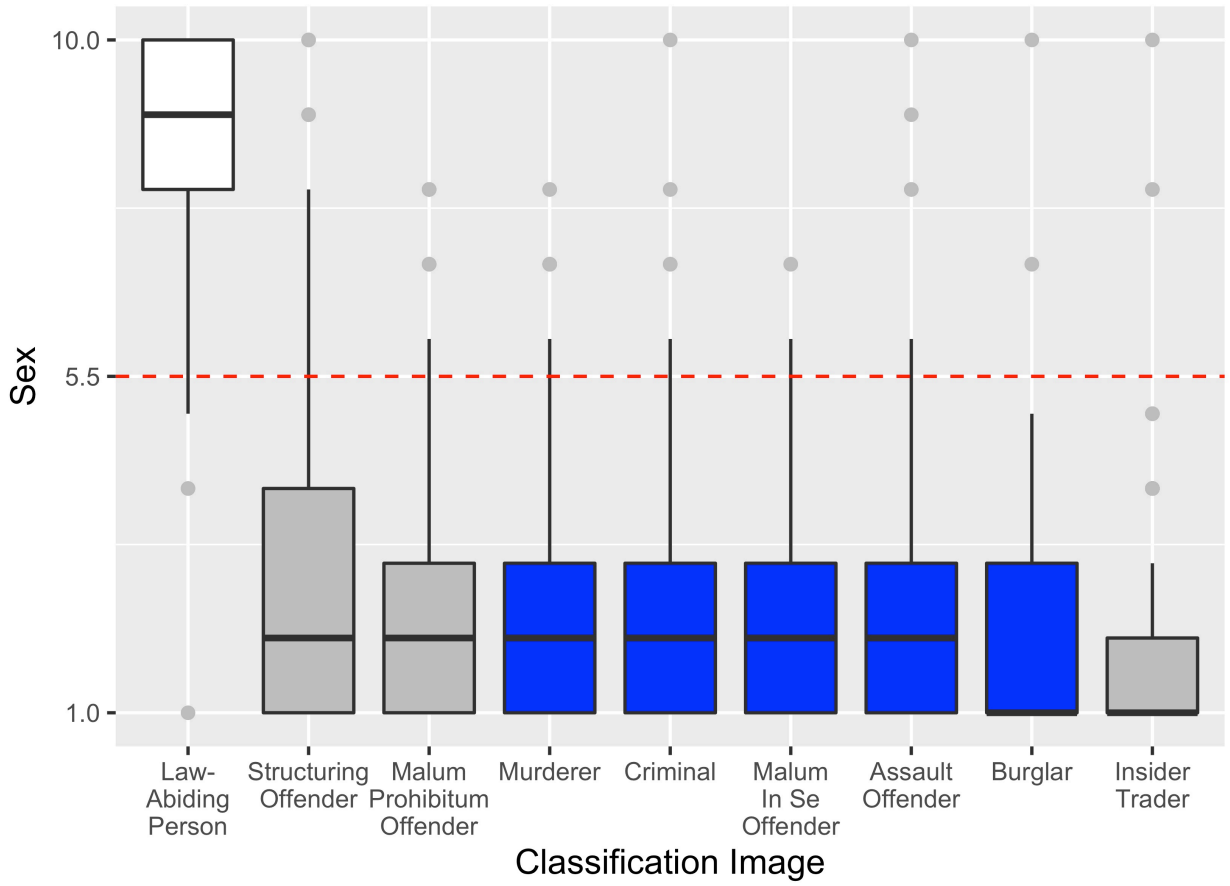


Figure 4.3 Male/female sex ratings for stereotype classification images. On the y-axis, 10 = definitely female and 1 = definitely male. The horizontal dotted line marks the midpoint of the scale. The colors align with the clusters presented in Figure 4.5.

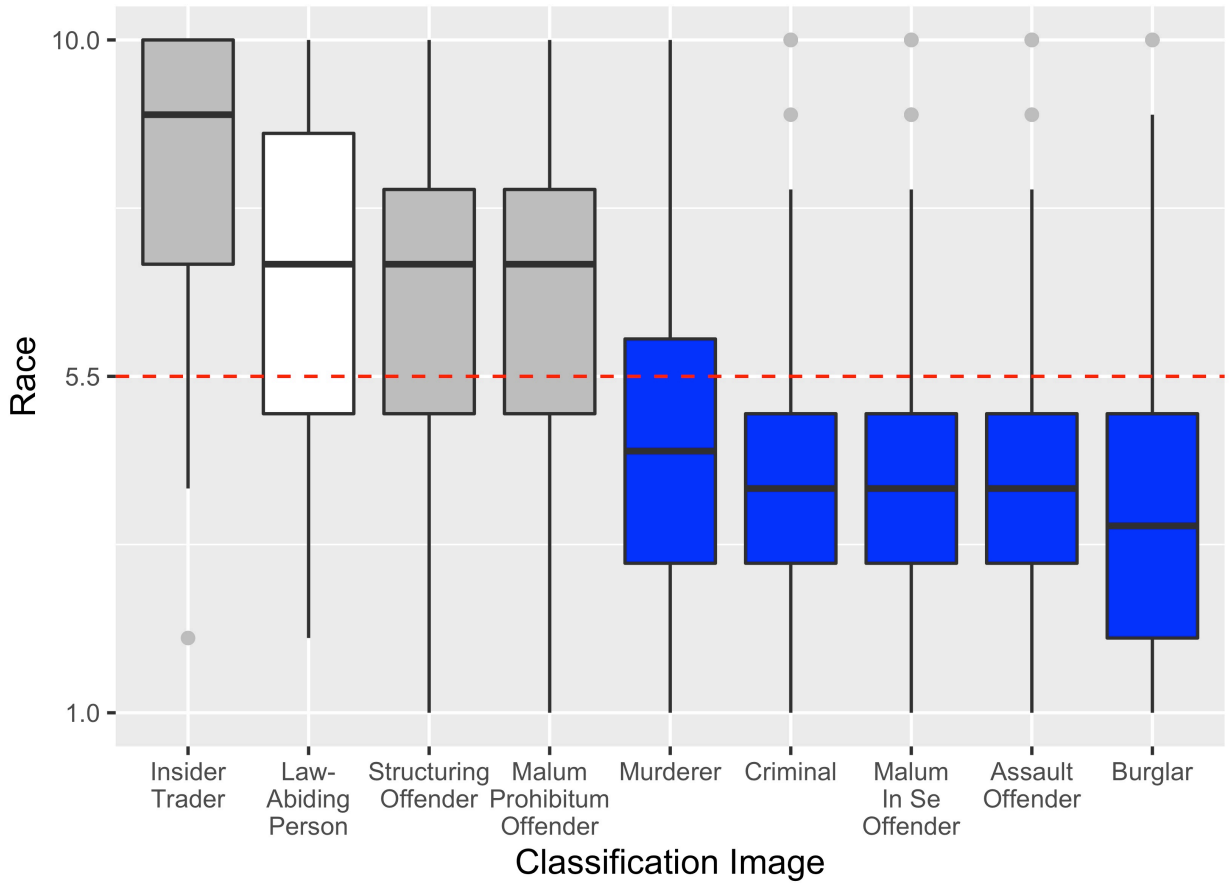


Figure 4.4 Black/White race ratings for stereotype classification images. On the y-axis, 10 = definitely White and 1 = definitely Black. The horizontal dotted line marks the midpoint of the scale. The colors align with the clusters presented in Figure 4.5.

Given these two sets of ratings, we tested whether they clustered into our three hypothesized groups: White female for law-abiding, White male for *malum prohibitum*, and Black male for *malum in se*. To determine the optimal number of clusters, we used silhouette analysis (Rousseeuw, 1987), which is a measure of how similar an object is to its own cluster as compared to other clusters, and to cluster the data we used k-means clustering, a machine learning algorithm that partitions observations into clusters such that each observation belongs to the cluster with the nearest mean center (Malik & Tuckfield, 2019).

Figure 4.5 shows the results of the clustering procedure, which yielded results that almost perfectly accorded with our hypothesis. The algorithm divided the data into three clusters, which might be thought of as White female, White male, and Black male clusters. Of primary importance are the *malum in se* and *malum prohibitum* results, which show opposing patterns (see the Appendix for the breakdown by percentage). Both of these classification images were assigned almost exclusively to the male clusters (Clusters 2 and 3), which accords with the association of criminality with maleness. Further, the *malum prohibitum* classification image was predominantly assigned to the White male cluster (Cluster 2), while the *malum in se* classification images was predominantly assigned to the Black male cluster (Cluster 3), and this pattern held for the other criminal subtypes that fall within these respective categories.

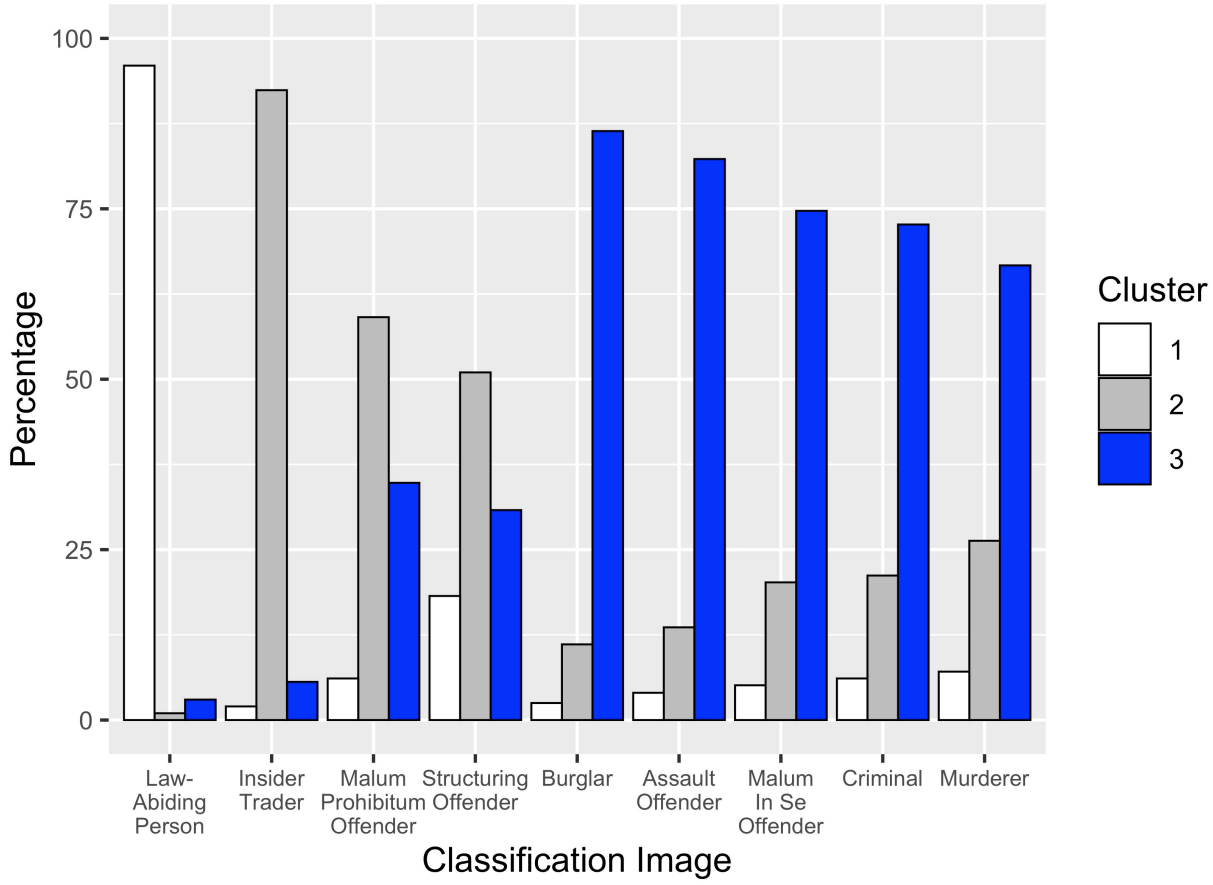


Figure 4.5 Results of k-means clustering of ratings of stereotype classification images. Cluster 1 can be thought of as a White female cluster, while Cluster 2 is a White male cluster, and Cluster 3 is a Black male cluster.

In summary, using an entirely data-driven approach, we generated images of criminal subtype stereotypes and clustered those images into their mathematically determined groups. The groups provided strong evidence of an association between crime and males, and, in particular, between crime and Black males. More importantly, they also showed the predicted nuance: not all criminals were imagined as Black males; some were imagined as White males, and this parsing between White and Black appeared to be related to the moral dimension, to the *malum prohibitum/in se* distinction.

Discussion

The results of our studies help us to understand an underlying dimension that people use to form stereotypical judgments of criminals. *Malum in se* crimes are associated with Black perpetrators. Crimes that are seen as immoral are seen as being committed by Black individuals. White individuals commit crimes that are technically illegal but not necessarily immoral. This is true whether we name specific crimes or just invoke the general category of morality. *Malum in se* crimes are perceived as committed by Black individuals, and *malum prohibitum* crimes are perceived as committed by White individuals.

Prior research has shown a relationship between race and criminality such that Black individuals were more likely to be associated with crimes than White individuals (Eberhardt, 2019; Mitchell et al., 2005). Research on criminal subtypes refined that conclusion to show that not all crimes were stereotypically associated with Black individuals (e.g., see Skorinko &

Spellman, 2013). However, prior research has not identified an underlying dimension that differentiated the crimes that were associated with these races. The current work extends prior research by showing that the moral dimension is at least one of the ways that people judge the racial prevalence of crimes.

Another interesting aspect of our results is the sex stereotype of crimes. Regardless of the type of crime we presented, the mental image that participants formed (in Study 4.1) was always rated as male by participants in Study 4.2. The only face judged to be female was that resulting from the law-abiding condition. This is consistent with other work using explicit rating scales (Esqueda, 1997; Gordon, Michels, & Nelson, 1996; Skorinko & Spellman, 2013; Sunnafrank & Fontes, 1983), but represents the first documentation of the sex stereotype using the reverse correlation approach.

Stereotypes matter in the real world of criminal justice. In studies of jurors in courtrooms and mock jurors in the laboratory, there is a tendency to disadvantage defendants who have characteristics matching those of the stereotypical offender (Sunnafrank & Fontes, 1983). The race-crime congruency effect is the most commonly studied variant of crime congruency, and its primary supposition, within a dichotomized Black/White population, is that Black defendants accused of stereotypically Black crimes will be disadvantaged, and White defendants accused of stereotypically White crimes will be disadvantaged. For example, Jones and Kaplan (2003) found that participants perceived Black men accused of auto theft as more guilty than White men so accused, and they perceived White men accused of embezzlement as more guilty than Black men so accused. Moreover, Petsko and Bodenhausen (2019) found that the race congruency effect was more pronounced for Black defendants than for Whites.

Limitations

While we provided evidence that stereotypes of immorality impact criminal subtype stereotypes, we did not attempt to unpack the precise magnitude of the contribution. For instance, when it comes to criminal subtype stereotype formation, what are the relative impact of base rates, individual experience, moral stereotypes, and so on? Future experimental work should explore designs that can quantify this.

Another limitation of the present chapter is that we focused on stereotypes in terms of Black and White. While this was in line with much past research (see Eberhardt, 2019) and reflects attitudes relating to two key racial groups within the United States, future work should explore stereotypes pertaining to other races and ethnicities, individuals with different skin tones, as well as other demographic factors. We know that people perceive others in light of their varied, numerous, and intersecting social identities, yet we know little about how intersectional person perception operates in the domain of legal decision-making (Mitchell, Haw, Pfeifer, & Meissner, 2005).

In addition, we believe reverse correlation research on criminal subtype stereotypes should be expanded to different and representative participant groups. While our student population was diverse and also was limited to our group of interest (American citizens), future work could explore how these stereotypes differ by age group, country, and other factors.

Future Directions

While a number of future directions emerge from this chapter, I would like to emphasize one in particular: media images of criminal offenders. News media have much latitude in deciding which photos to publish and disseminate, and some research has indicated that, in

making these decisions, they reinforce general stereotypes of Black criminality (Colburn & Melander, 2018). Within this general trend, I might explore whether news outlets are more likely to include a photo of a criminal that fits with the subtype expectation. That is, Black individuals who commit crimes in the *malum in se* category are more likely to have their photos included than are White individuals who commit crimes in the *malum prohibitum* category. This imbalance, in speaking to the underlying moral stereotype identified in this chapter, would shed much light on what motivates the appeal (in terms of "clicks") and perpetuation (in terms of "shares") of such harmful disseminations.

Conclusion

Can we know what our stereotypes of criminal subtypes actually look like? In this chapter, we show that, indeed, we can, and with much accuracy. Our use of the reverse correlation method confirmed the female/male distinction in stereotypes regarding who is generally law-abiding and who is generally criminal. Moreover, the method confirmed various criminal subtype stereotypes that have been identified in the past survey research: that, for example, assault is a stereotypically Black offense and insider trading a stereotypically White offense (Esqueda, 1997; Gordon, Michels, & Nelson, 1996; Skorinko & Spellman, 2013; Sunnafrank & Fontes, 1983). Given the social desirability concerns that adversely affect stereotype research, the fact that we were able to replicate these stereotypes using an entirely data-driven approach shows that this method is worth continued and expanded use.

More importantly, the method uncovered a distinct difference in how *malum prohibitum* and *malum in se* defendants are imagined, with the former being stereotyped as White and the latter as Black. While criminal subtype stereotypes are certainly influenced by crime base rates,

by individuals' experiences, and by a multitude of other factors, we found that criminals who commit crimes that are perceived as not as immoral are imagined as White individuals, while those offenders who commit crimes that seem more immoral are imagined as Black individuals.

From this foundation, I move to Chapter 5, where I explore whether and how these imputations of immorality impact case outcomes.

Note

The original research presented in this chapter was conducted by Joel Cooper, DongWon Oh, and the dissertation author.

Chapter 5

How Imputations of Immorality Color Criminal Subtype Stereotypes and Impact Case Outcomes

In Chapter 5, I unpack how the morality-driven stereotypes discussed in Chapter 4 help explain an inconsistency that has plagued research regarding the effects of stereotypes on case outcomes; specifically, research regarding crime congruency. The race-crime congruency effect is often found with stereotypically Black crimes but not with stereotypically White crimes. I provide evidence that Black defendants are stereotyped as immoral actors, and this has a dramatic effect on case outcomes. For guilt determinations, this imputation of immorality overwhelms the crime congruency effect, thus leading to the observed Black/White imbalance. In short, the theory helps to explain racially disparate rates of both guilt adjudication and sentence impositions.

Criminal Subtypes and the Crime Congruency Effect

Criminal subtype stereotypes have important consequences for decisions in the courtroom, and these consequences are typically considered through the lens of the *crime congruency effect*. At its most fundamental, the effect refers to the tendency, for a given crime, to disadvantage defendants who have characteristics matching those of the stereotypical offender (Sunnafrank & Fontes, 1983).

In crime congruency studies, outcomes can be divided into two groups. Evidentiary outcomes include recall of case facts and, of course, guilt. Thus, the crime congruency hypothesis is that when a defendant matches the racial stereotype for the crime, he or she will be perceived as more likely to have committed the crime, and participants will recall more case facts, especially the race of the defendant. There also are non-evidentiary matters, such as

interpersonal regard (will participants think less of stereotypical defendants?) and sentence imposition (will participants punish stereotypical defendants more severely?).

Race-crime congruency research developed rapidly in the 1980s (see Bodenhausen, 1988; Bodenhausen & Wyer, 1985). However, there have been perplexing limitations to the findings. Gordon and colleagues (1988) found the predicted interaction of race by crime type (burglary vs. embezzlement), but the crime-specific effects were there only for the stereotypically Black crime. In other words, when the crime was the stereotypically White offense of embezzlement, the crime congruency effects disappeared. Similarly, in a 1993 study, Gordon found that participants perceived a Black defendant accused of assault as more guilty than a White assault defendant, but a White embezzler was not perceived as more guilty than a Black embezzler.

The inconsistencies have persisted. As noted in Chapter 4, Skorinko and Spellman (2013) reported the results of two studies on the race-crime congruency effect. In the first, they found no crime congruency effects for guilt or for sentencing decisions. The only effect found was that, when the defendant's race was crime congruent, there was better recall of the defendant's race. In the second study, there was the same memory effect, and guilt and sentencing outcomes were inconclusive.

Given these inconsistences, Bodenhausen and Petsko (2019) revisited the race-crime congruency effect. Using a more highly powered sample of participants from Amazon's Mechanical Turk, they were able to replicate earlier work showing a bias against Black defendants in the case of stereotypically Black crimes. Importantly, they failed to find a corresponding bias against White defendants in the case of stereotypically White crimes. Moreover, when the outcome measure asked what participants personally would decide (rather than what they believed the average American juror would decide), the outcomes were always

harsher towards White defendants, which led the authors to speculate that these results likely were "contaminated by social desirability concerns" (p. 77).

As discussed in Chapter 4, we found a distinct difference in how *malum prohibitum* and *malum in se* defendants are imagined, with the former being stereotyped as White and the latter as Black. While criminal subtype stereotypes are certainly influenced by crime base rates, by individuals' experiences, and by a multitude of other factors, we found that criminals who commit crimes that are perceived as not as immoral are imagined as White individuals, while those offenders who commit crimes that seem more immoral are imagined as Black individuals. Building on these results, in the present chapter we predict that these morality-driven stereotypes help to explain the inconsistencies that have long plagued the crime congruency literature.

Guilt vs. Sentencing

There is one other aspect of crime congruency that is worth discussing. While past research has largely lumped the different outcomes (evidentiary and non-evidentiary ones) together, it may be a mistake to consider these outcomes as reflective of similar processes. For example, for the most common evidentiary-focused outcome—guilt—the analysis is straightforward: can you say that, beyond a reasonable doubt, the defendant did what he or she is alleged to have done? It makes sense that crime congruency would impact such an analysis, since, in the absence of countervailing evidence, participants might view the stereotype as probative of specific criminal propensity. This fully accords with the social scientific understanding of how stereotypes impact judgments, such that stereotypes are used as heuristics for interpreting behavior (Dipboye, Fromkin, & Wiback, 1975; Grant & Holmes, 1981), and activation of a stereotype may lead individuals to interpret information in accordance with the stereotype (Duncan, 1976; Sagar & Schofield, 1980).

However, for the most frequent and important non-evidentiary-focused outcome—sentencing—the analysis might be quite different, predicated on more complex sets of issues. In other words, sentencing involves a different set of issues that may or may not be related to our tendency to infer culpability. What punishment should a defendant receive? The answer may hinge on how one conceives of the purpose of criminal punishment. As Weinhofen (1929), Alschuler (2003), and Booth (2019) have shown, all of the following questions might be asked. Is it primarily about rehabilitation? Or is it about incapacitation? Or perhaps specific deterrence, general deterrence, retribution, restitution, normative validation—or even something else (Carlsmith, Darley, & Robinson, 2002)? Of these, the only ones that are clearly relevant to crime congruency are those that may pertain to recidivism. That is, a participant might deem a stereotypical defendant as more likely to commit that same crime again in the future. Thus, incapacitation and specific deterrence might be relevant to crime congruency. It is not at all clear that the other justifications are relevant. In light of this, we predict that crime congruency will be impactful for guilt determinations, but when it comes to sentencing, its effect will be much diminished.

The Plan of the Research

Across three studies, we used a novel paradigm to examine the impact of the moral distinction on judgments of guilt and innocence, as well on the imposition of sentencing.

The Experiments

Our reverse correlation procedure suggested that when a crime roughly fits into the *malum in se* category, such criminals are expected to be Black males. When a crime roughly fits into the *malum prohibitum* category, such criminals are expected to be White males. There is an obvious question that proceeds from this finding: in what ways does it matter? In the criminal

justice context, there are numerous ways in which it might matter. For instance, if police expect a Black individual to be the perpetrator of a specific crime, then they might be more likely to interrogate Black individuals. Indeed, there are numerous reports of such profiling concerns (Snook, Eastwood, Gendreau, Goggin, & Cullen, 2007). Once an arrest is made, there are two primary areas that warrant consideration: guilt adjudication and sentence imposition. Do these criminal stereotypes impact determinations of guilt, and do they impact subsequent sentencing decisions?

As discussed in the introduction, these questions are often considered through the lens of crime congruency. However, crime congruency, and especially race-crime congruency, has yielded perplexing results: the effect appears consistent for stereotypically Black crimes, but much less consistent for stereotypically White crimes. Moreover, there are difficulties in experimentally testing matters having to do with race and the law. Social desirability concerns may keep people from participating with candor and openness (Krumpal, 2013).

Thus, the most objective way to add to our scientific understanding of these basic processes and avoid unnecessary complications, is to test crime congruency theory using populations about which participants have little to no knowledge. More specifically, we would want to disregard race, gender, and other defendant characteristics, create new group stereotypes that parse along the observed moral fault line (*malum prohibitum/in se*), and use a jurisdiction in which the defendants are completely novel to our participants.

Researchers working on moral intuition have overcome similar difficulties by embedding their scenarios in alien environments (Rottman & Kelemen, 2012). Following this lead, we set our studies in an alien environment. Specifically, we described a group of Martians, noted that

they could be identified by their physical characteristics (light green scales and purple eyes), and established the group stereotypes that we sought to explore.

With this design, we were able to test the hypotheses laid out in the introduction to this article. Building on Chapter 4, we operated with the assumption that criminal subtype stereotypes partly reflect deeper stereotypes about the moral behavior of different racial groups. In the current studies, we hypothesized that crime congruency would be relevant when the outcome measure was guilt-related; that is, people would count a defendant's proximity to a specific crime stereotype as evidence of the likelihood that the defendant committed the crime. Further, we predicted that moral stereotyping also would be at play, and it would override the crime congruency effect in predictable ways: if a group was stereotyped as generally moral, then that perceived morality would counteract the effect of the criminal stereotype. Finally, for non-evidentiary decisions such as sentencing, crime congruency might play a role, but a smaller one, as the outcome would depend largely on the imputations of morality/immorality, which is more germane to the purposes of punishment, as discussed above.

Study 5.1

In Study 5.1, we tested whether the crime congruency effect was evident within the confines of our alien world.

Methods. Participants read a description of a crime stereotype, where the crime was either one of two that are stereotypically considered Black (assault or auto theft) or one of two offenses stereotypically considered White (embezzlement or tax evasion), in our actual world, of course. But we exported these stereotypes to our faraway planet. For example, in the embezzlement condition, the description was as follows: "Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. One notable

group of Martians are called the Namoz: they have light green scales and purple eyes. For some reason, when Martians think of embezzlement, they think of Namoz. If one were to imagine a Namoz committing a crime, the general thinking is that the crime would be something like embezzling $500 from one's employer." The other three conditions were identical except for the crime type and its description (i.e., assault, auto theft, or tax evasion).[13]

The crime descriptions were selected from the National Survey of Crime Severity so that the two stereotypically Black crimes were relatively equal in perceived severity, and the two stereotypically White crimes were likewise equal in perceived severity (Wolfgang et al., 1985).[14] Next, the crime descriptions were rated by participants recruited through an online survey service for perceived morality/immorality. No differences were observed between the two stereotypically Black crimes, nor were any observed between the two stereotypically White crimes (for both, $p > .85$ and Cohen's $d < .02$).[15] In other words, the stereotypically Black crimes can be considered equally severe and equally immoral, and the same for the stereotypically White crimes.

After viewing one of the prompts, participants stated their expectations for the defendant in terms of either guilt or sentence.[16] For the guilt condition (pre-conviction), regardless of whether a participant saw the assault or auto theft condition (stereotypically Black crimes), the participant was told that the defendant was a Namoz, had recently been arrested for assault, and had pleaded not guilty. Similarly, regardless of whether a participant saw the embezzlement or tax evasion condition (stereotypically White crimes), the participant was told that the defendant

[13] See the Appendix for all four prompts.
[14] The crimes selected were slightly revised to better match, and this revision was done prior to testing on perceived morality/immorality.
[15] See the Appendix for additional analyses regarding these ratings.
[16] See the Appendix for all of the guilt and sentence prompts.

was a Namoz, had recently been arrested for embezzlement, and had pleaded not guilty. The dependent variable was the question, "How do you think a Martian jury will rule on whether the defendant committed the crime?" The question was rated on a 6-point Likert scale (0 = definitely not guilty; 5 = definitely guilty).[17]

For the sentence condition (post-conviction), regardless of whether a participant saw the assault or auto theft condition (stereotypically Black crimes), the participant was told that the defendant was a Namoz, had recently been convicted of assault, and had entered the sentencing phase of the process. Similarly, regardless of whether a participant saw the embezzlement or tax evasion condition (stereotypically White crimes), the participant was told that the defendant was a Namoz, had recently been convicted of embezzlement, and had entered the sentencing phase of the process. The dependent variable was the question, "How do you think a Martian judge will rule on how severely the defendant should be punished?" The question was rated on a 6-point Likert scale (0 = minimum required by law; 5 = maximum required by law).

In summary, the design was such that the crime was either stereotypically Black or stereotypically White, and the defendant was either crime congruent or incongruent; and there were two conditions: participants decided either guilt or sentence.

Lastly, participants completed questions regarding their demographics and their qualifications for inclusion (such as whether they were U.S. citizens).

A total of 400 participants were recruited through Prolific, 100 for each of the four types: pre-conviction for stereotypically White crimes, post-conviction for stereotypically White crimes, pre-conviction for stereotypically Black crimes, post-conviction for stereotypically Black

[17] Following the method used in previous research, we phrased the outcomes so that participants opined on how others would decide, as such psychological distance is thought to lead to more truthful answers, ones that more closely reflect actual behavior (see Petsko & Bodenhausen, 2019).

crimes. After exclusions for attention check or eligibility failure, 382 remained. All were United States citizens who were living in the United States. Forty-nine percent were male, 49% were female, and the remaining selected transgender, other, or preferred not to disclose. Mean age (± SD) of the participants was 33 ± 13 years. The racial breakdown was 74% White/Caucasian, 7% Black/African American, 6% Hispanic/Latinx, 8% Asian/Pacific Islander, 1% Native American, 1% Middle Eastern/Arab, 3% selected "other," and the remaining participants preferred not to disclose.

Results. For the stereotypically White crimes, there was a significant difference in guilt adjudication by defendant congruency ($M = 4.06$, $SE = .09$) and incongruency ($M = 3.72$, $SE = .12$): $t(96) = 2.20$, $p = .03$, Cohen's $d = .44$, such that congruent defendants were more likely to be found guilty. However, for sentence severity, the crime congruency effect was not observed: there was not a significant difference in sentence severity by defendant congruency ($M = 2.83$, $SE = .22$) and incongruency ($M = 2.60$, $SE = .22$): $t(93) = .72$, $p = .47$, Cohen's $d = .15$.[18]

For the stereotypically Black crimes, the same results were observed. There was a significant difference in guilt adjudication by defendant congruency ($M = 4.27$, $SE = .10$) and incongruency ($M = 3.80$, $SE = .11$): $t(93) = 3.14$, $p = .002$, Cohen's $d = .64$, such that congruent defendants were more likely to be found guilty. However, for sentence severity, the crime congruency effect was not observed: there was not a significant difference in sentence severity by defendant congruency ($M = 3.23$, $SE = .20$) and incongruency ($M = 3.02$, $SE = .18$): $t(92) = .76$, $p = .45$, Cohen's $d = .16$.[19]

[18] See the Appendix for related figures.
[19] See the Appendix for related figures.

Overall, for both stereotypically Black crimes and stereotypically White crimes, crime congruent defendants were deemed more likely to be guilty, even on a remote planet. Once found guilty, however, crime congruent defendants were not necessarily expected to receive more severe punishment.

Study 5.1's design lacked two elements that we sought to address in the succeeding studies. First, the stereotypes were related only to the crime types and did not include a moral dimension; this moral dimension is added in Studies 5.2 and 5.3. Second, Study 5.1's design compared only same-race stereotypical crimes (assault vs. auto theft; embezzlement vs. tax evasion). Study 5.3 will explore crime congruency across race stereotypical categories (*malum prohibitum* vs. *malum in se*).

Study 5.2

In Study 5.2, as a first assay into the relationship between crime congruency and imputations of group morality, we introduced stereotypes of morality and immorality. We limited ourselves to general criminality, rather than the criminal subtypes, which we explore in Study 5.3. In short, we created a stereotype for the Martian group that reflected either positive moral associations or more negative moral associations, that is, associations with immorality.

Methods. Study 5.2's procedure was similar to that used in Study 5.1. Participants (i) read a description of a group stereotype, and (ii) either read a description of either an arrest and were tasked with deciding guilt, or read a description of an arrest with a subsequent conviction and were tasked with deciding the sentence.

The group stereotype was either moral/law-abiding or immoral/criminal. The moral/law-abiding stereotype introduced the stereotype like this: "[O]ne group of Martians are called the Namoz. As soon as you see a Namoz, with their light green scales and their purple eyes, you

know they belong to a group of good, moral Martians. It's not so easy to imagine a Namoz committing a crime." The immoral/criminal stereotype had a few key differences: "As soon as you see a Namoz, with their light green scales and their purple eyes, you know they belong to a group of bad, immoral Martians. It's easy to imagine a Namoz committing a crime."[20] Note that, as in Study 5.1, we linked the stereotype to appearance, thus mirroring stereotypes that attach to White and Black individuals.

After viewing one of the prompts, participants stated their expectations for the defendant in terms of either guilt or sentence. For the guilt condition (pre-conviction), the prompt and dependent variable were as follows. Regardless of whether a participant saw the moral or immoral group stereotype, the participant was told that the defendant was a Namoz, had recently been arrested for a crime, and was claiming innocence. The dependent variable was the question, "How do you think a Martian jury will rule on whether the defendant committed the crime?" The question was rated on a 6-point Likert scale (0 = definitely not guilty; 5 = definitely guilty).

For the sentence condition (post-conviction), regardless of whether a participant saw the moral or immoral group stereotype, the participant was told that the defendant was a Namoz, had recently been convicted of a crime, and had entered the sentencing phase of the process. The dependent variable was the question, "How do you think a Martian judge will rule on how severely the defendant should be punished?" The question was rated on a 6-point Likert scale (0 = minimum required by law; 5 = maximum required by law).

Lastly, participants completed questions regarding their demographics and their qualifications for inclusion (such as whether they were U.S. citizens).

[20] See the Appendix for all materials.

A total of 200 participants were recruited through Prolific, 100 for the guilt variant (pre-conviction) and 100 for the sentence variant (post-conviction). After exclusions for attention check or eligibility failure, 196 remained. All were United States citizens who were living in the United States. Forty-seven percent were male, 51% were female, and the rest selected "other" or preferred not to disclose. Mean age (± SD) of the participants was 37 ± 15 years. The racial breakdown was 83% White/Caucasian, 4% Black/African American, 3% Hispanic/Latinx, 8% Asian/Pacific Islander, <1% Middle Eastern/Arab, 2% selected "other," and <1% of participants preferred not to disclose.

Results. For both guilt (pre-conviction) and sentence (post-conviction), there was a significant difference across the groups, such that congruent defendants were deemed more likely to be found guilty and more likely to be sentenced harshly. For guilt adjudication, we observed a significant difference by defendant congruency ($M = 3.70$, $SE = .11$) and incongruency ($M = 2.58$, $SE = .16$): $t(95) = 5.80$, $p < .001$, Cohen's $d = 1.18$, such that crime congruent defendants were considered more likely to be guilty. Likewise, for sentence severity, crime congruent defendants were considered likely to receive harsher punishment; there was a significant difference by defendant congruency ($M = 3.04$, $SE = .20$) and incongruency ($M = .78$, $SE = .15$): $t(97) = 8.95$, $p < .001$, Cohen's $d = 1.80$.[21]

Overall, for general criminality (i.e., the crime type was not specified), crime congruent (vs. incongruent) defendants were deemed more likely to be guilty and, once found guilty, they were expected to receive more severe punishment. However, with Study 5.2's design, we could not disentangle the effects of the moral/immoral stereotypes and the crime congruency effect. That is, both the guilt and sentence outcomes might be resultant of moral stereotypes, crime

[21] See the Appendix for related figures.

congruency, or both, given that the congruent group was also the stereotypically immoral group. This is addressed in Study 5.3, where the design allows us to vary congruency across the moral/immoral stereotypes.

Study 5.3

In Study 5.3, we were able to test the key elements of our hypotheses: although crime congruency is impactful when the outcome measure is guilt-related, moral stereotyping also is impactful, and it should override the crime congruency effect in predictable ways. Specifically, if a group is stereotyped as generally moral, then that perceived morality should counteract the effect of the criminal stereotype. Finally, for non-evidentiary decisions, such as sentencing, crime congruency should play only a small role, if any, as the outcome largely should depend upon the moral/immoral group stereotypes.

Methods. Study 5.3's procedure was similar to that used in Studies 5.1 and 5.2, only now the design was a 2 x 2, where group stereotype (moral/*malum prohibitum* offender vs. immoral/*malum in se* offender) and criminal type (crime congruent vs. crime incongruent) were varied. As before, participants either participated in a guilt (pre-conviction) or a sentence (post-conviction) version.[22]

In sum, the study consisted of the following: participants (i) read a description of a group stereotype (moral/*malum prohibitum* offender vs. immoral/*malum in se* offender); (ii) read a description of an arrest (or an arrest and conviction) for either a *malum prohibitum* or a *malum in se* offense; (iii) decided guilt or sentence.

For the moral/*malum prohibitum* stereotype, as in Studies 5.1 and 5.2, participants were told about a group of Martians, called the Namoz, who were distinguishable by their physical

[22] See the Appendix for all materials.

appearance. The Namoz were known as a group of generally good Martians, and if one were to imagine a Namoz committing a crime, the general thinking was that the crime would be something like a *malum prohibitum* crime. *Malum prohibitum* was then defined and examples were given.

For the immoral/*malum in se* stereotype, participants likewise were told about a group of Martians, called the Namoz, who were distinguishable by their physical appearance. The Namoz were known as a group of generally bad Martians, and if one were to imagine a Namoz committing a crime, the general thinking was that the crime would be something like a *malum in se* crime. *Malum in se* was then defined and examples were given.

Next, participants were told about an alleged or convicted criminal, and were asked their expectations for the defendant in terms of either guilt or sentence. For the guilt condition (pre-conviction), the prompt and dependent variable were as follows. In one condition, participants were told that the defendant, a Namoz, had been arrested for a *malum prohibitum* offense. (Thus, in this condition, the moral/*malum prohibitum* group stereotype represented the crime congruent defendant.) Specifically, they were told that the defendant had been arrested for a structuring offense. In order to make it clear that it was a *malum prohibitum* offense, participants were told that it was not clear that anyone had been harmed in any way by the crime. Moreover, participants were told that, while there was some evidence that the defendant had committed the crime, it was not a sure victory for the prosecution.

In the other condition, participants were told that the defendant, a Namoz, had been arrested for a *malum in se* offense. (Thus, in this condition, the immoral/*malum in se* group stereotype represented the crime congruent defendant.) Specifically, they were told that the defendant had been arrested for assault and battery. In order to make it clear that it was a *malum*

in se offense, participants were told that the victim had suffered injuries on account of the crime. Moreover, participants were told that, while there was some evidence that the defendant had committed the crime, it was not a sure victory for the prosecution.

For both conditions, the dependent variable was the question, "How do you think a Martian jury will rule on whether the defendant committed the crime?" The question was rated on a 6-point Likert scale (0 = definitely not guilty; 5 = definitely guilty).

For the sentence condition (post-conviction), the prompt and dependent variable were identical, except now participants were told that the defendant had been convicted of the crime, and the case was entering the sentencing phase. The dependent variable was the question, "How do you think a Martian judge will rule on how severely the defendant should be punished?" The question was rated on a 6-point Likert scale (0 = minimum required by law; 5 = maximum required by law).

Lastly, participants completed questions regarding their demographics and their qualifications for inclusion (such as whether they were U.S. citizens).

For the guilt (pre-conviction) variant, a total of 200 participants were recruited through Prolific. After exclusions for attention check or eligibility failure, 191 remained. All were United States citizens who were living in the United States. Forty-six percent were male, 53% were female, and less than 1% selected transgender or preferred not to disclose. Mean age (± SD) of the participants was 34 ± 13 years. The racial breakdown was 72% White/Caucasian, 5% Black/African American, 10% Hispanic/Latinx, 10% Asian/Pacific Islander, <1% Middle Eastern/Arab, 2% selected "other," and <1% of participants preferred not to disclose.

For the sentence (post-conviction) variant, a total of 200 participants were recruited through Prolific. After exclusions for attention check or eligibility failure, 184 remained. All

were United States citizens who were living in the United States. Forty-two percent were male, 56% were female, 2% transgender, and the rest selected "other." Mean age (± SD) of the participants was 34 ± 13 years. The racial breakdown was 73% White/Caucasian, 5% Black/African American, 9% Hispanic/Latinx, 8% Asian/Pacific Islander, <1% Native American, 1% Middle Eastern/Arab, and 4% selected "other."

Results. A two-way between-subjects ANOVA revealed a significant effect for the interaction of criminal type (congruent vs. incongruent) and group stereotype (moral vs. immoral) on guilt adjudication: $F(3, 187) = 8.61$, $p = .004$, $\eta^2 = .04$. Post-hoc comparisons using two-tailed t-tests and the Tukey HSD test yielded the following. For the *malum prohibitum* offense, there was virtually no difference in guilt adjudication by defendant congruency ($M = 3.35$, $SE = .14$) and incongruency ($M = 3.48$, $SE = .14$): $t(95) = .66$, $p = .51$, $p_{adj} = .91$, Cohen's $d = .13$, such that congruent defendants were deemed no more likely to be found guilty (see Figure 5.1). However, for the *malum in se* offense, there was a significant difference in guilt adjudication by defendant congruency ($M = 3.88$, $SE = .09$) and incongruency ($M = 3.19$, $SE = .16$): $t(92) = 3.55$, $p < .001$, $p_{adj} = .004$, Cohen's $d = .74$, such that congruent defendants were deemed more likely to be found guilty (see Figure 5.1). In summary, the crime congruency effect was observed, but only when the crime congruent defendant came from the stereotypically immoral group (right half of Figure 5.1). By contrast, when the crime congruent defendant was not from the stereotypically immoral group, the crime congruency effect disappeared (left half of Figure 5.1).

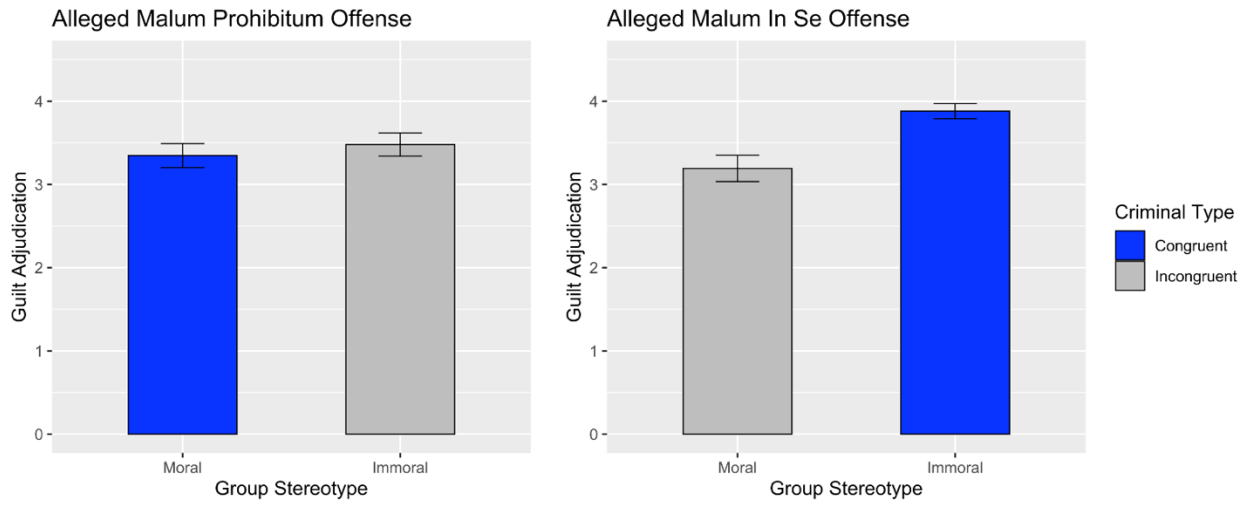


Figure 5.1 Guilt adjudication for crime congruent and incongruent defendants accused of either a *malum prohibitum* or *malum in se* offense.

For the sentence (post-conviction) variant, a two-way between-subjects ANOVA revealed a significant effect for the interaction of crime type (congruent vs. incongruent) and group stereotype (moral vs. immoral) on sentence severity: $F(3, 180) = 47.02$, $p < .001$, $\eta^2 = .18$. Post-hoc comparisons using two-tailed t-tests and the Tukey HSD test yielded the following. For the *malum prohibitum* offense, there was a significant difference in sentence severity by defendant congruency ($M = .91$, $SE = .21$) and incongruency ($M = 2.39$, $SE = .24$): $t(90) = 4.70$, $p < .001$, $p_{adj} < .001$, Cohen's $d = .98$, such that congruent defendants were punished less severely (see Figure 5.2). For the *malum in se* offense, there was a significant difference in sentence severity by defendant congruency ($M = 3.78$, $SE = .15$) and incongruency ($M = 2.32$, $SE = .23$): $t(90) = 5.00$, $p < .001$, $p_{adj} < .001$, Cohen's $d = 1.07$, such that congruent defendants were punished more severely (see Figure 5.2). In summary, the moral/immoral group stereotype appeared to overrun any crime congruency effects, such that, regardless of congruency, the defendant from the stereotypically immoral group was punished more severely.

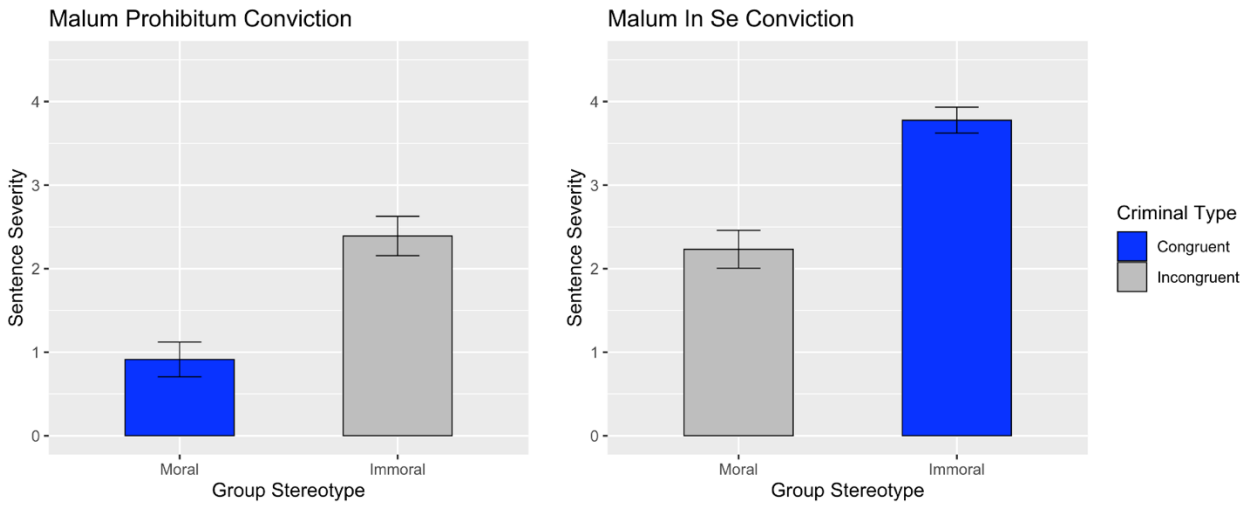


Figure 5.2 Sentence severity for crime congruent and incongruent defendants convicted of either a *malum prohibitum* or *malum in se* offense.

Overall, the results of Study 5.3 show that, for guilt determinations, crime congruency mattered only when the congruent group's stereotype suggested immorality. That is, when the group stereotype for a crime congruent defendant suggested immorality, the defendant was deemed more likely to be guilty. No effect was detected when the group stereotype for a congruent defendant suggested morality. In addition, once found guilty, crime congruency did not appear to matter. Regardless of whether the defendant was a crime congruent or crime incongruent criminal, if the defendant's group stereotype suggested immorality, the defendant was deemed more likely to be severely punished.

Summary of Experimental Results

If we bring the results of our Martian studies back to Earth, what is most striking is that we were able to replicate the particulars of the body of research on crime congruency. In short, we were able to replicate the imbalance in the crime congruency effect, that it is evident for stereotypically Black crimes but not as consistently evident with stereotypically White crimes. Our use of the Martian Namoz population revealed that this imbalance can be generated in the

absence of White and Black individuals. That is, in an environment without Black and White races, when we retained what we believe to be the content of the underlying stereotype—that is, when we created moral and immoral stereotypes for the Martians—the crime congruency effect was identical to that found on Earth with stereotypically White and Black crimes: it operates for Black defendants accused of stereotypically Black crimes but not for White defendants accused of stereotypically White crimes. This suggests that the moral component, which we identified in our reverse correlation work, carries into case outcomes and may be a precipitating force behind the crime congruency effect.

General Discussion

Building off of our finding that *mala prohibita* crimes are associated with White individuals and *mala in se* crimes with Black individuals, we reached a better understanding of the race-crime congruency effect. First, we confirmed that crime congruency yields a consistent, although small, impact on guilt determinations. When the evidence was ambiguous, the criminal subtype stereotypes we created impacted guilt. For example, when the defendant came from a group that was associated with embezzlement, that defendant was deemed more likely to be found guilty of embezzlement. The same was true for assault.

However, sentencing was a somewhat distinct matter, and thus we reiterate our assertion that, for non-guilt outcomes, such as sentencing, the crime congruency effect lacks explanatory power. We did not find crime congruency effects for sentencing in our Studies 5.1 and 5.3. This makes sense, given that sentencing decisions rely on a host of factors and intuitions besides those relating to criminal propensity.

Our primary hypothesis was that, in addition to crime congruency, moral stereotyping would be at play and would override the crime congruency effect in predictable ways. In line

with this, we found that when a crime congruent defendant came from a group that was stereotyped as generally moral, the perceived morality counteracted the effect of the criminal stereotype in guilt determinations. This finding helps to explain a series of results that have vexed crime congruency scholars for decades: crime congruency may be a real effect, but it often is not impactful on guilt determinations when the crime is a stereotypically White crime (Gordon, 1993; Gordon, Bindrim, McNicholas, & Walden, 1988; Petsko & Bodenhausen, 2019). The present research shows that this peculiarity likely is owing to the added impact of stereotypes of morality/immorality. In short, and as a correlate to the above description, when the crime incongruent defendant is Black, the perceived immorality of the defendant overrides the incongruency, and the defendant is found as guilty or more guilty than a White crime congruent defendant.

Conclusion

In Chapter 5, I shed light on how the morality-driven stereotypes identified in Chapter 4 help to explain why the race-crime congruency effect is often found with stereotypically Black crimes but not with stereotypically White crimes. Because Black defendants are simultaneously stereotyped as immoral actors, this imputation of immorality overwhelms the crime congruency effect, thus leading to the observed Black/White imbalance. These results represent a significant advance in our understanding of criminal subtype stereotypes and their impact, and they set the stage for much future work in the field.

Note

The original research presented in this chapter was conducted by Joel Cooper and the dissertation author.

Part III

Can Legal Technology Be Used to Decrease Bias in the Law?

Part I and II of this dissertation dealt with the problem: bias in the law. Part III deals with solutions and, especially, technology-driven ones. When faced with evidence of racially biased behavior or racially disparate outcomes, the solution *de rigueur* in many domains is to institute implicit bias trainings: Starbucks did as much after a well-publicized incident of bias occurred in one of its stores (Noguchi, 2018); Michael Bloomberg, in his bid for the U.S. presidency, touted the trainings for physicians (Sullivan, 2019). Such trainings also are common when bias in the law is found: the California State Assembly is set to vote on three bills that would require implicit bias training for law enforcement officers, judges, and trial attorneys (Eberhardt & Banks, 2019). This response is an admirable one, as it brings attention to issues of racial bias, and it also expresses a belief that humans can change, that attitudes are mutable.

But do such interventions work? In this part of my dissertation, I survey the literature regarding bias interventions in the law, which suggests that such interventions have only limited efficacy. Of those that show promise, the change in implicit bias is generally small, and any subsequent change in behavior is negligible (Forscher et al., 2019). Long-term success is most important, and here there is scant evidence that interventions work at all (Chapman et al., 2018; Devine, Forscher, Austin, & Cox, 2012; Lai, Hoffman, & Nosek, 2013; Payne, Vuletich, & Lundberg, 2017; Sawyer & Gampa, 2018). Lai and colleagues (2016) reviewed nine interventions that had shown short-term efficacy, finding that for none of the nine interventions did the positive effects persist for more than a few days. Part of the reason that racial bias is so hard to eradicate is that it is automatic and deeply ingrained, representing a sorting process that humans habitually conduct. Stereotypes result from a need for coherence, simplicity, and

predictability (Tajfel, 1981), and they provide a basis for explaining behavior (Bodenhausen & Wyer, 1985; Kahneman & Tversky, 1972).

In light of these concerns, in Part III of this dissertation I argue that, within the law, the most promising route forward is to guide human discretion through specially developed artificial intelligence tools. In other words, I argue that we should artfully and skillfully decrease human discretion by building systems that provide real-time statistics and flag potential instances of bias—and which can be readily incorporated into decisional workflows.

In Chapter 7, I outline the approach through two examples, one in criminal and one in civil law. First, I present a method and process for how prosecutors' offices can use their own historical data (replete with biases) to build computational models that output racially-unbiased outcomes. These models can be used to anchor human decision making. Then, I discuss the development of an artificial legal intelligence that predicts the outcomes of civil legal cases, and I outline how it might be used to decrease bias in civil legal outcomes.

But before I broach these solutions, I explore—in Chapter 6, a potential reason why such systems have not been used to-date, a reason that stands to inhibit their adoption moving forward. Specifically, I explore perceptions of legal technology and, more specifically, how bias in legal technology is perceived and assessed. I ask, is there reverse automation bias for legal technology? If there is, does it threaten the adoption of technology that holds promise for decreasing racial bias in legal outcomes? I show that, given biased outcomes, machine adjudicators are trusted less than human adjudicators. Moreover, I show that this attitudinal backlash has potentially grave consequences: machine decision makers that are as adept at legal decision making tasks and more racially fair than human decision makers are still trusted less than the human decision makers.

Chapter 6

Attitudes Regarding Bias in Legal Technology

Decades of research has taught us that the implementation of technology can have significant benefits for decision making tasks. More than 65 years ago, Meehl (1954) reviewed 20 studies, including ones in the legal domain, found that machines forecast better than humans. Similarly, Dawes showed that this was true even when the machine was merely a simple linear model or a model with equal weights (Dawes, 1979; Dawes, Faust, & Meehl, 1989). Grove, Zald, Lebow, Snitz, and Nelson (2000) reported the results of a meta-analysis of over a hundred studies that investigated expert predictions, finding that algorithms significantly outperformed human forecasters (see also Ægisdóttier, 2006; Silver, 2012). Thus, it should not be surprising to read this strong conclusion from Gottfredson and Gottfredson, written in 1986: "In virtually every decision making situation for which the issue has been studied, it has been found that statistically developed predictive devices outperform human judgments… This is one of the best-established facts in the decision-making literature" (p. 247).

In addition, scholars are increasingly finding that machine decision making has significant potential for decreasing social and racial disparities. Kleinberg and colleagues (2018) conducted groundbreaking research showing that machine predictions could lead to dramatically improved legal outcomes. The authors found that their policy simulation would reduce crime by up to 25% without causing an increase in jail populations. Alternatively, it would be possible to keep crime rates steady while reducing jail populations by 42%. More importantly, the authors found that these gains could be achieved while simultaneously reducing racial disparities. In his qualitative review of the literature, Miller (2018) found that algorithms produce less biased (in terms of racial, gender, and other prominent biases) outcomes than human decision makers—and

this was true in surprising applications, including ones ranging from mortgage lending (Gates, Perry, & Zorn, 2002) to selection of corporate board members (Erel, Stern, Tan, & Weisbach, 2018). In addition, there are various innovative ways that machines can be used to overcome human biases (for example, see Avery & Cooper, 2020b, 2020c; Gecker, 2019).

Resistance to Legal Technology

In light of this, it might be surprising to learn that there is overwhelming resistance to the adoption of legal technology, and this resistance is largely motivated by claims that such technology will lead to racially biased outcomes. Four recent books have challenged technology on this ground (Benjamin, 2019; Eubanks, 2017; O'Neil, 2016; Pasquale, 2016). In addition, entire organizations are devoted to this stance. The AI Now Institute at New York University focuses on identifying bias in artificial intelligence (AI) and its impact on human rights (Whittaker et al., 2018). The Institute has partnered with the American Civil Liberties Union, publishing a number of papers expressing wariness of the increasing reliance on technology in the United States.

While there are many valid reasons to be skeptical of machine solutions, this push against legal technology is particularly surprising given that it will necessitate continued reliance on human legal actors. In the United States, racial disparities have characterized the legal system since the nation's founding and especially since the end of the U.S. Civil War, and such racial disparities are a human problem: they are driven by the human-designed legal system, which is implemented and conducted by humans, most notably defense attorneys, prosecutors, judges, and jurors (Munger & Seron, 2017). If machines show potential for operating as or more fairly than these human actors, it would make sense to take steps towards at least including machine

decision making in the law. But the resistance to legal technology remains strong, and we seem more concerned with bias in legal technology than with bias in human legal actors.

Reverse Automation Bias

Why would there be extreme distrust of machines when they show such promise, especially for decreasing bias? One reason may be found in a quirk of attitudes towards machines: when machines fail, there often is an extreme attitudinal backlash against them. In fact, research has shown that machines generally benefit from "automation bias," such that information they present is viewed as inherently trustworthy (Cummings, 2004). However, should they fail even once, this benefit evaporates (Lee & Moray, 1992, 1994; Parasuraman & Manzey, 2010). This phenomenon—sometimes called the "first-failure effect"—is persistent: recovery in trust may occur, but it occurs at a very slow rate (Merlo, Wickens, & Yeh, 2000; Moray, 2003; Moray & Inagaki, 2000). Perhaps the strongest evidence of reverse automation bias comes from Dietvorst, Simmons, and Massey (2015), who call the effect "algorithm aversion." They found that, even when an algorithm showed predictive power that surpassed that of a human decision maker, if the algorithm made any mistakes, then observers would show greater aversion to the algorithm (vs. the human) decision maker.

The public has witnessed this effect with modern technology failures. For example, consider that, in the United States, 102 people die per day as the result of automobile accidents, the vast majority of which are caused by human driver error (NHTSA, 2017). However, when a pedestrian died as the result of an accident with a self-driving Uber vehicle (Meyer, 2018), there was public outcry that has not waned in the time that has passed since the accident. Such backlash also is evident in public opinion regarding companies like Facebook, which once were praised but now are scorned given perceived breaches of trust that have occurred (Romm, 2018).

Likewise, there have been highly publicized technology failures in various public domains, including healthcare (Obermeyer, Power, Vogeli, & Mullainathan, 2019). Most apposite to this chapter, there have been well-publicized instances of potentially machine generated bias within the law, including bias in predictive policing (Rieland, 2018) and in criminal risk assessments (Angwin, Larson, Mattu, & Kirchner, 2016; Barocas & Selbst, 2016).

Current Research

In this chapter, I explore this conundrum at the intersection of technology, legal decision making, and racial bias by asking three questions. First, is there reverse automation bias for legal technology? Second, if there is, does it threaten the adoption of technology that shows promise for decreasing racial bias in legal outcomes? Lastly, what factors are driving the reverse automation bias?

The research progresses as follows. In the first two studies, I seek to establish whether there is reverse legal automation bias by exploring trust in human and machine legal decision makers. The machine is identified as an algorithm trained to automate prosecutors' decision making in plea bargaining. I explore participants' trust ratings for these adjudicators (the algorithm and the prosecutor) when racial disparities are discovered, and we also explore participants' response in terms of trying to correct for the bias. I hypothesize that, given equal rates of bias in human and machine adjudicators, the machines will be trusted less. In addition, more drastic corrective action will be recommended in response to the machine bias. In the final set of studies, I investigate whether there are differences in trust when the adjudicators are established as equally adept at the decision making task. In particular, I stipulate parity in all intangibles across the human and machine adjudicators, thus allowing us to isolate the specific effects of racial bias on trust. I also test an extreme difference—twice as much bias in humans

vs. machines—to see the extent to which humans are "let off the hook" while the machines are castigated. I again hypothesize that the machines will be trusted less, even when they are significantly more equitable than the human decision makers.

In total, these studies allow formulation of an answer to the question of whether machines are being held to a standard more demanding than what is asked of human decision makers. Are lay individuals' preferences in legal adjudication leading them to undervalue machines—and is this undervaluation at the expense of racial equality?

Study 6.1

In Study 6.1, I tested our hypothesis that lay individuals have more trust in human (vs. machine) legal adjudicators.

Method

Participants. A total of 105 participants were recruited through Prolific, with approximately half pre-screened for White race and half pre-screened for Black race. After exclusions, 102 remained. All were United States citizens who were living in the United States. Forty-three percent were male, 56% were female, and 1% were gender variant/non-conforming. Mean age (± SD) of the participants was 32 ± 11 years. The racial breakdown was 51% White/Caucasian and 49% Black/African American.

Procedure. Participants read a description of the criminal justice process in a fictional U.S. county. (All materials for the experiments presented in this chapter are contained in the Appendix.) They were told that nearly every criminal case in the county resolves by plea bargaining. When a defendant is arrested, the prosecutor decides which offenses with which to charge the defendant. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached.

In the human condition, participants were told that the process is entirely human-driven. Prosecutors generally have an idea as to what case outcomes should be, and they negotiate with these outcomes in mind. In the machine condition, participants were told that there is an algorithm that has been built to determine case outcomes. This machine (the algorithm) tells the prosecutor what the outcome should be, and the prosecutor negotiates with this machine-generated outcome in mind. Finally, in both conditions, participants were told that the prosecutors/the algorithm went through an extensive training program designed to ensure fair and just outcomes.

The dependent variable was a single item related to trust. "To what extent do you agree or disagree with the following statement? I trust the performance of the prosecutors [machine] in Cooke County." The question was rated on a 6-point Likert scale that ranged from 0 (No, definitely disagree) to 5 (Yes, definitely agree).

Lastly, participants completed questions regarding their demographics.

Results

When asked how much they trusted the adjudicator, participants were more likely to trust the human prosecutor ($M = 3.14$, $SE = .15$) than the machine ($M = 2.10$, $SE = .18$). A two-tailed t-test revealed that the difference in trust was significant, $t(100) = 4.43$, $p < .001$, Cohen's $d = .88$ (Figure 6.1).

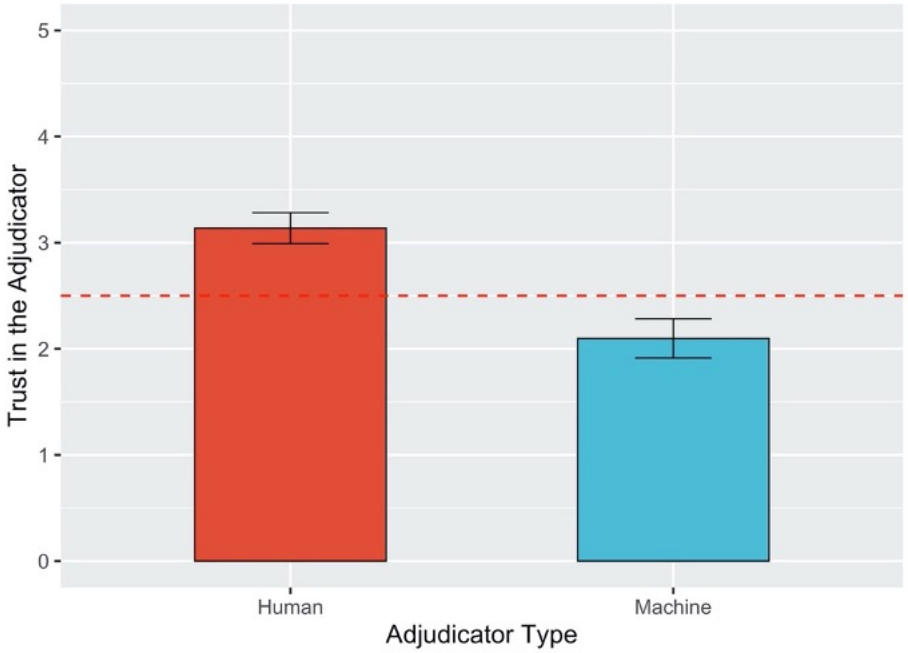


Figure 6.1 Trust in the legal adjudicator (human vs. machine). The horizontal dotted line represents the midpoint of the scale, which also represents the midpoint between indicating that one trusts (scores of 3-5) or does not trust (scores of 0-2) the adjudicator.

As an exploratory analysis, I considered attitudes held by Black and White participants. Both racial groups had significantly more trust in the human adjudicator than the machine adjudicator. Race appeared not to play a role. A two-way ANOVA showed no significant interaction between participants' race and adjudicator type, nor was there any effect of race on trust in the human or machine adjudicator.

Summary of Study 6.1

Lay individuals had more trust in humans than machines for a legal decision making task, and this was equally true for White and Black participants. In addition, on average, participants not only had more trust in humans, but they trusted (in absolute terms) human adjudicators and distrusted machine adjudicators.

Study 6.2

In Study 6.2, I tested whether the lack of trust in machines we identified in Study 6.1 holds consequences for racial equality. More specifically, I hypothesized that lay individuals would be more permissive of racial bias that arises from human, rather than machine, legal decision making.

Method

Participants. A total of 105 participants were recruited through Prolific, with approximately half pre-screened for White race and half pre-screened for Black race. After exclusions, 101 remained. All were United States citizens who were living in the United States. Forty-two percent were male, 57% were female, and 1% were transgender male. Mean age (± SD) of the participants was 31 ± 13 years. The racial breakdown was 48% White/Caucasian and 52% Black/African American.

Procedure. As in Study 6.1, participants read a description of the criminal justice process in a fictional U.S. county. They were told that nearly every criminal case in the county resolves by plea bargaining.

In the human condition, participants were told that the process is entirely human-driven. Prosecutors run the process and determine the final plea bargain the defendant receives. In the machine condition, participants were told that there is an algorithm that has been built to determine case outcomes. This algorithm (an AI machine interface) runs the process and determines the final plea bargain the defendant receives. Finally, in both conditions, participants were told that the prosecutors/the algorithm went through an extensive training program designed to ensure fair and just outcomes.

In both conditions, participants were told that, although the evidence was preliminary, it appeared that the prosecutor/the algorithm was giving White defendants plea bargains that were approximately 7% better than those given to similarly-situated Black defendants.

The dependent variable was designed to reflect how serious participants considered the disparities and how committed they would be to seeing that the disparities were addressed. "There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (they would have to postpone everything) until these disparities are eliminated?" The question was rated on a 6-point Likert scale, ranging from 0 (definitely no) to 5 (definitely yes).

Lastly, participants completed questions regarding their demographics.

Results

The results revealed that participants were more in favor of shutting down the office until the racial disparities are addressed when the adjudicator was a machine ($M = 3.84$, $SE = .19$) rather than human ($M = 2.94$, $SE = .21$): $t(99) = 3.13$, $p = .002$, Cohen's $d = .62$ (Figure 6.2). In other words, participants wanted to take more drastic corrective measures when the bias was driven by a machine adjudicator.

As an exploratory analysis, I considered attitudes held by Black and White participants. A two-way ANOVA did not reveal any significant interaction between race and adjudicator type, although the results did show, as one might expect, that Black participants were more in favor of taking the corrective measures (M = 3.77) than were White participants (M = 2.94).

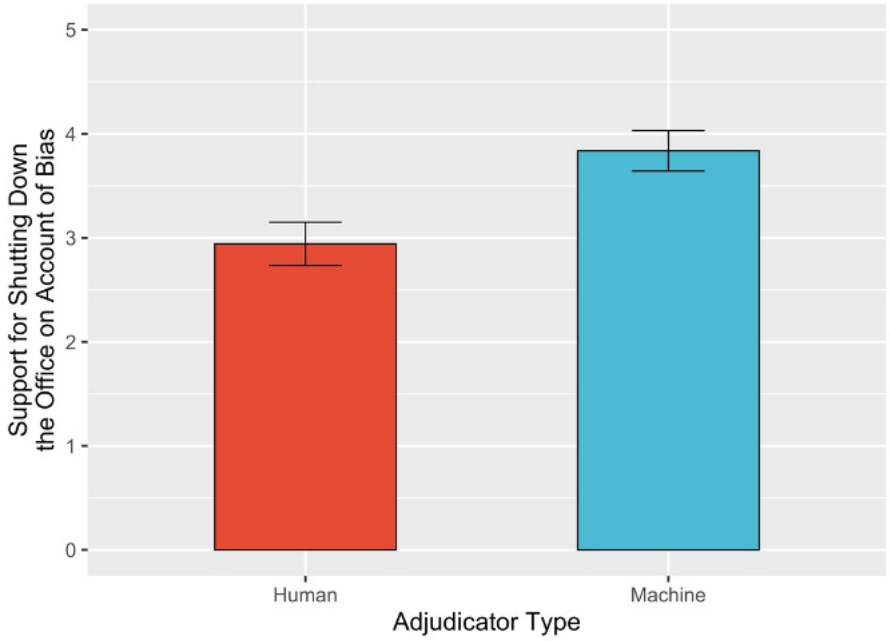


Figure 6.2 Professed support for shutting down legal operations until racial bias in the adjudicators (human vs. machine) is addressed and corrected. Greater scores indicate greater support for shutting down the office.

Summary of Study 6.2

When racial disparities arose in legal decision making, participants were more likely to believe that drastic corrective action was called for when the disparities were generated by a machine adjudicator as opposed to a human adjudicator. In short, machines were being held to a more demanding standard. In addition, I found significant main effects of participant race, such that Black participants considered racial bias—whether by machines or humans—as more serious and more worthy of intensive corrective action.

Study 6.3

In Study 6.3, I built upon the previous studies in two primary ways. First, I wanted to see whether recognition of fallibility (of bias in the adjudicators) would affect trust. That is, if both human and machine legal adjudicators are showing bias, is there less trust in one of them? We

hypothesized that machine bias would be more detrimental to trust scores, and the machines would be trusted less. Second, I wanted to expand the measure of bias response. In addition to the drastic measure of shutting down the office, I wanted to see if even a more anodyne measure—increasing funding to investigate the bias—would be disparately supported. That is, I hypothesized that participants would support funding to investigate machine legal bias to a greater extent than they would support funding to investigate human legal bias.

Method

Participants. A total of 205 participants were recruited through Prolific. After exclusions, 195 remained. All were United States citizens who were living in the United States. Fifty-two percent were male, 45% were female, <1% were transgender male, and 3% selected "other." Mean age (± SD) of the participants was 32 ± 11 years. The racial breakdown was 46% White/Caucasian and 54% Black/African American.

Procedure. As in Studies 6.1 and 6.2, participants read a description of the criminal justice process in a fictional U.S. county. They were told that nearly every criminal case in the county resolves by plea bargaining.

In the human condition, participants were told that the process is entirely human-driven. Prosecutors run the process and determine the final plea bargain the defendant receives. In the machine condition, participants were told that there is an algorithm that has been built to determine case outcomes. This algorithm determines the final plea bargain the defendant receives. In both conditions, participants were told that, even though the prosecutors/the algorithm went through an extensive training program designed to ensure fair and just outcomes, they/it are not perfect and do show some bias in favor of White defendants. A recent review of

the office found that White defendants were receiving plea bargains that were approximately 7% better than those given to similarly-situated Black defendants.

There were three dependent variables. First, as in Study 6.1, participants indicated the extent to which they agreed or disagreed with the following statement: "I trust the performance of the prosecutors [machine (the algorithm)] in Cooke County." The question was rated on a 6-point Likert scale, ranging from 0 (No, definitely disagree) to 5 (Yes, definitely agree).

Next, participants completed the following two items, which were presented in random order, counterbalanced across the total number of participants. As in Study 6.2, these items were designed to reflect the how serious participants considered the disparities and how committed they would be to seeing that the disparities were addressed. One item was: "There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (they would have to postpone everything) until these disparities are eliminated?" The question was rated on a 6-point Likert scale, ranging from 0 (definitely no) to 5 (definitely yes).

The other item was: "The office does not have any funds budgeted to investigating this issue. However, it could shift funding away from other worthy causes in order to pursue this. How much funding should be redeployed so that this issue can be investigated further?" The question was rated on a 6-point Likert scale, ranging from 0 (None) to 5 (As much as possible).

Lastly, participants completed questions regarding their demographics.

Results

A two-tailed t-test revealed a significant difference in trust by adjudicator type. Confirming results of the prior study, we found that participants trusted the human adjudicators (M = 1.78, SE = .12) more than they trusted the machine adjudicator (M = 1.31, SE = .11): t(193)

= 2.77, p = .006, Cohen's d = .40 (alternatively, one might say that they distrusted the humans less than they distrusted the machine). In short, given equal rates of White preference bias, participants retained more trust in the biased humans than in the biased machines.

For the two measures of interventions (How much funding should be reallocated to investigating the issue? Should the entire office be shut down until the bias is addressed?), we again found attitudes that were consistently harsher towards machine (vs. human) bias. A two-tailed t-test revealed a significant difference, by adjudicator type, in support for reallocating funds to address the disparities, such that participants supported greater reallocation when the adjudicator was machine (M = 3.29, SE = .12) rather than human (M = 2.94, SE = .12): t(193) = 2.11, p = .04, Cohen's d = .30. Likewise, with the more drastic measure, a two-tailed t-test revealed a significant difference, by adjudicator type, in support for shutting down the office until the racial disparities were addressed; participants professed greater support for shutting down the office when the adjudicator was machine (M = 3.30, SE = .13) rather than human (M = 2.80, SE = .14): t(193) = 2.64, p = .009, Cohen's d = .38.

As in the first two studies, there was no interaction based on race: not for the trust outcome, the funding decision, or the decision to shut down the office. However, we found significant main effects of participant race, such that Black participants had less trust overall (M = 1.78 vs. M = 1.34) and were more likely to support greater funding for investigating the issue (M = 3.33 vs. M = 2.87), but they were not more likely to want to shut down the office.

Summary of Study 6.3

In Study 6.3, I found that when human and machine legal adjudicators showed equal rates of racial bias, there was less trust in the machine adjudicator. In addition, regardless of whether a proposed corrective measure was reasonable (increase funding to investigate) or extreme (shut

down the office), participants were more in favor of it when the biased adjudicator was a machine. In short, there was greater commitment to eradicating bias in machines than in humans.

Studies 6.4 and 6.5

Across the previous studies, I found consistent evidence that people trust machine legal adjudicators less than they trust human legal adjudicators. This was true in general, and it was true when participants were told about equal rates of racial bias in the machine and human adjudicators. I also found that, given equal rates of bias, people were motivated to investigate and even shut down the machines, while they were much more permissive of the humans' bias.

These results prompted some important follow-up questions. First, could racial bias have equal effects on trust, but because people trust machines less overall (irrespective of racial bias), it appears that machine bias is being taken more seriously? In Studies 6.4a, 6.4b, and 6.4c, I address this. Specifically, I stipulate parity in all intangibles across the human and machine adjudicators, thus allowing me to isolate the specific effects of racial bias on trust in them. I also tested an extreme difference—twice as much bias in humans vs. machines—to see the extent to which humans are "let off the hook" and the machines illogically disparaged. In Study 6.5, I take steps towards explaining what might be driving the trust ratings that we have found.

Study 6.4a

In Study 6.4a, I established parity across the human and machine adjudicators, specifying that, during its training, the algorithmic prosecutor was not only designed to ensure fair and just outcomes, but it also was shown to perform as well as the average prosecutor. The idea was to ensure that there was equal trust in the machine and human adjudicators before we tested attitudes about bias in these adjudicators in the succeeding studies.

Method

Participants. A total of 200 participants were recruited through Prolific, with approximately half pre-screened for White race and half pre-screened for Black race. After exclusions, 192 remained. All were United States citizens who were living in the United States. Forty-seven percent were male, 47% were female, 2% were transgender male, 3% were gender variant/non-conforming, and <1% selected "other/not listed." Mean age (± SD) of the participants was 32 ± 13 years. The racial breakdown was 52% White/Caucasian and 48% Black/African American.

Procedure. As in the first two studies, participants read a description of the criminal justice process in a fictional U.S. county. They were told that nearly every criminal case in the county resolves by plea bargaining.

In the human condition, participants were told that the process is entirely human-driven. Prosecutors run the process and determine the final plea bargain the defendant receives. Participants were told that new prosecutors in Cooke County go through an extensive training program designed to ensure fair and just outcomes, and this program, which is run in collaboration with a local university, is considered state-of-the-art.

In the machine condition, participants were told that there is an algorithm that has been built to determine case outcomes. This algorithm determines the final plea bargain the defendant receives. Participants were told that the machine (the algorithm) was created through an extensive training program designed to ensure fair and just outcomes, and this program, which was run in collaboration with a local university, is considered state-of-the-art. In addition, participants were told that, after training, the machine was tested against the performance of actual prosecutors in the office, and its performance was equal to the average performance of the

prosecutors. “In other words, this is a machine that is as good at being a prosecutor as an average prosecutor is.”

The dependent variable was the same trust measure that we used in the earlier studies. Participants indicated the extent to which they agreed or disagreed with the following statement: “I trust the performance of the prosecutors [machine (the algorithm)] in Cooke County.” The question was rated on a 6-point Likert scale, ranging from 0 (No, definitely disagree) to 5 (Yes, definitely agree). After the trust item, there were nine additional items, which we analyze and discuss in Study 6.5. Lastly, participants completed questions regarding their demographics.

Results

A two-tailed t-test failed to reveal a significant difference in trust by adjudicator type, such that a significant difference between trust in human adjudicators (M = 2.70, SE = .11) and trust in the machine adjudicator (M = 2.50, SE = .14) was not found: $t(190) = 1.13$, $p = .26$, Cohen’s $d = .16$. As in the first three studies, there was no interaction based on race: a two-way between-subjects ANOVA failed to reveal a significant effect for the interaction of race (Black vs. White participants) and adjudicator type (human vs. machine) on trust.

In summary, this new design appeared successful. After reading that the machine had been tested and was shown to perform as well as the average human prosecutor, trust ratings for human vs. machine adjudicators were roughly equivalent.

Study 6.4b

In Study 6.4b, building from the same prompts as in the Study 6.4a, I introduced a finding of bias. In particular, I told participants that White preference bias had been found in the adjudicator (machine or human), and the specific level of bias described was equal in both conditions. Given the results of Study 6.4a, I could be relatively confident that I was now testing

our hypothesis that, given parity across human and machine adjudicators, including parity in bias, the impact of reported bias on attitudes regarding these adjudicators would be most severe when the adjudicator was a machine.

Method

Participants. A total of 150 participants were recruited through Prolific. After exclusions, 149 remained. All were United States citizens who were living in the United States. Forty-seven percent were male, 46% were female, 1% were transgender male, 4% were gender variant/non-conforming, and 1% preferred not to answer. Mean age (± SD) of the participants was 31 ± 12 years. The racial breakdown was 65% White/Caucasian, 3% Black/African American, 9% Hispanic/Latinx, 17% Asian/Pacific Islander, <1% Native American, <1% Middle Eastern/Arab, 4% Multiracial, and <1% preferred not to answer.

Procedure. The procedure was identical to Study 6.4a, except participants were now told that, even though the prosecutors/the algorithm went through an extensive training program designed to ensure fair and just outcomes, they/it are not perfect and do show some bias in favor of White defendants. A recent review of the office found that White defendants were receiving plea bargains that were approximately 6% better than those given to similarly-situated Black defendants.

As in Study 6.4a, I included the trust measure. Participants indicated the extent to which they agreed or disagreed with the following statement: I trust the performance of the prosecutors/the machine (the algorithm) in Cooke County. The question was rated on a 6-point Likert scale, ranging from 0 (No, definitely disagree) to 5 (Yes, definitely agree). I also included a measure designed to reflect the how serious participants considered the disparities and how committed they would be to seeing that the disparities were addressed. “There has been some

discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (they would have to postpone everything) until these disparities are eliminated?" The question was rated on a 6-point Likert scale, ranging from 0 (definitely no) to 5 (definitely yes).

Lastly, participants completed questions regarding their demographics.

Results

I found that participants trusted human adjudicators (M = 2.05, SE = .13) much more than they trusted the machine adjudicator (M = 1.52, SE = .16), with a two-tailed t-test revealing a significant difference: $t(147) = 2.51$, $p = .01$, Cohen's $d = .41$.

In addition, a two-tailed t-test revealed a significant difference, by adjudicator type, in support for shutting down the office until the racial disparities were addressed, such that participants professed greater support for shutting down the office when the adjudicator was a machine (M = 3.01, SE = .17) rather than human (M = 2.46, SE = .16): $t(147) = 2.39$, $p = .01$, Cohen's $d = .39$. Participants wanted to take more drastic corrective measures when the bias was driven by a machine adjudicator.

In summary, given equal rates of bias in machine and human legal adjudicators, participants had less trust in the machine adjudicator. In addition, they were much more likely to recommend shutting down the entire prosecutorial system until the bias was addressed when the bias was driven by the machine, rather than the human, adjudicator.

Study 6.4c

In Study 6.4c, I tested an extreme difference in bias: we reported to participants twice as much bias in the human (vs. machine) prosecutors. I hypothesized that, even given this extreme difference, attitudes and responses would be more positive towards the human adjudicators.

Method

Participants. A total of 150 participants were recruited through Prolific. After exclusions, 148 remained. All were United States citizens who were living in the United States. Fifty-two percent were male, 46% were female, <1% were transgender male, and 1% were gender variant/non-conforming. Mean age (± SD) of the participants was 30 ± 11 years. The racial breakdown was 55% White/Caucasian, 4% Black/African American, 8% Hispanic/Latinx, 20% Asian/Pacific Islander, <1% Native American, <1% Middle Eastern/Arab, 11% Multiracial, and <1% preferred not to answer.

Procedure. The procedure was identical to Study 6.4b, except participants in the human prosecutor condition were told that a recent review of the office found that White defendants were receiving plea bargains that were approximately 8% better than those given to similarly-situated Black defendants, and participants in the machine prosecutor condition were told that White defendants were receiving plea bargains that were approximately 4% better than those given to similarly-situated Black defendants. The measures were the same as those in Study 6.4b.

Results

The results showed greater trust in the human adjudicators (M = 2.09, SE = .12) than the machine adjudicator (M = 1.58, SE = .16), with a two-tailed t-test yielding $t(146) = 2.59$, $p = .01$, Cohen's d = .43 (Figure 6.3).

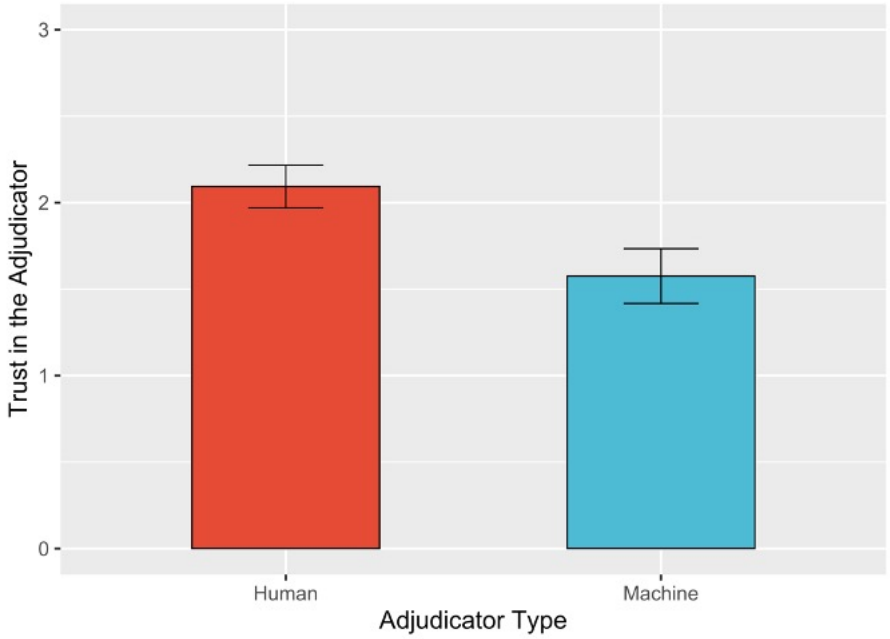


Figure 6.3 Trust in the legal adjudicator (human vs. machine) when the human adjudicator shows twice as much racial bias as the machine adjudicator.

The results showed greater support for shutting down the office until the racial disparities were addressed when the adjudicator was machine ($M = 3.10$, $SE = .16$) rather than human ($M = 2.33$, $SE = .15$), with a two-tailed t-test yielding $t(146) = 3.47$, $p < .001$, Cohen's $d = .57$. In other words, participants wanted to take more drastic corrective measures when the bias was driven by a machine adjudicator.

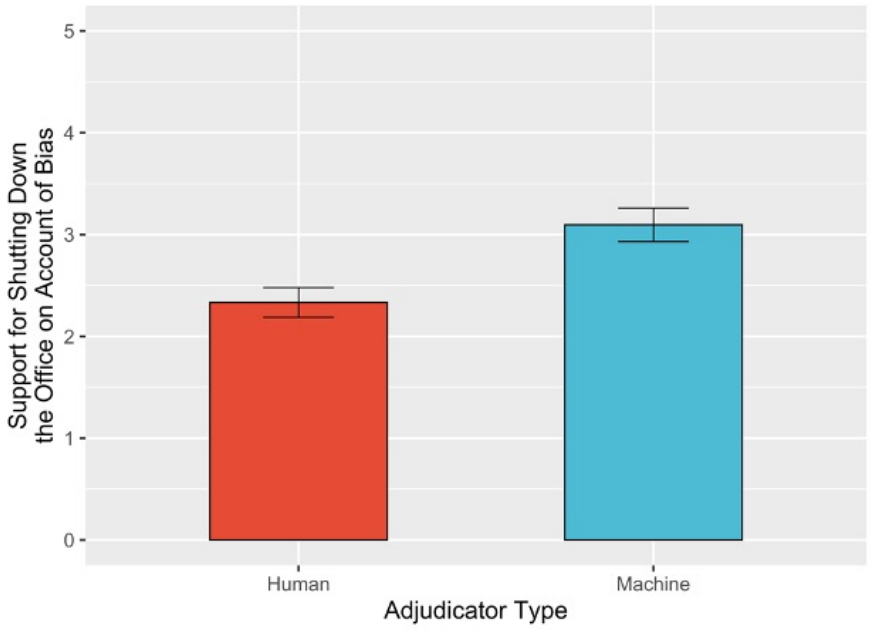


Figure 6.4 Professed support for shutting down legal operations until racial bias in the adjudicators (human vs. machine) is addressed and corrected. Greater scores indicate greater support for shutting down the office. Participants were informed that the human adjudicator showed White preference bias at rates double those of the machine adjudicator.

In summary, Study 6.4c revealed the extreme extent of individuals' reverse legal automation bias. When human legal adjudicators showed double the bias of machine legal adjudicators, participants still had less trust in the machines. Moreover, they were much more likely to recommend shutting down the entire prosecutorial system until the bias was addressed when the bias was driven by the machine, rather than the human, adjudicator. In short, it appeared that, whenever a machine showed bias, it would be discounted to a far greater extent than even a more biased human would be.

Study 6.5

There was one final issue that we sought to address. Through an exploratory analysis, I wanted to unpack what was possibly driving the trust ratings that we found in the preceding studies. First, legal decision making is often considered a moral undertaking, and people tend to

prefer that machine decision making is limited to mechanical tasks (Lee, 2018). At its most general, we might understand this concern as one relating to a general reluctance to allow machines to make important decisions (Dawes, 1979), but it is more specific than that. When a decision falls with the general category of a moral one, we tend to care how that decision was reached. Everett, Pizarro, and Crockett (2016) showed that how people arrive at their moral decisions influences whether those decisions are perceived as morally permissible. Decision makers were perceived as less trustworthy when their decisions were the result of merely calculating costs and benefits, rather than a deliberative struggle or a demonstrated sensitivity to others' welfare. For instance, one reason people don't trust autonomous vehicles is because the vehicles lack moral capacity (Young & Monroe, 2019).

Whereas humans possess highly complex minds capable of nuanced feelings, machines, including artificial intelligence, are perceived as entities that, while able to make decisions, lack nuance and depth of feeling (Gray, Gray, & Wegner, 2007). Weisman, Dweck, and Markman (2017) found that inferences of mindedness are more readily granted to decision makers who have the ability to value others' feelings and understand the moral ramifications of their behavior. Thus, if a machine decision maker is perceived as lacking a mind, people may view its decisions as morally deficient.

In addition to perceived lack of moral capacity, I hypothesized that individuals might fear that machine decision making runs the risk of "technological lock-in" (Crootof, 2019). Many perceive machines as unable to learn (Dawes, 1979) and prefer humans with their presumed ability to improve through experience (Highhouse, 2008).

Below, I introduce a few additional factors, including ones related to the notion that machines failed to properly consider individual targets and nuanced, qualitative data (Grove & Meehl, 1996).

Method

Participants and Procedure. Study 6.5 made use of the same design, population, and data as Study 6.4a. I deemed Study 6.4a the appropriate design for exploring what was driving trust perceptions because it was designed for parity across the decision makers. Participants were told that, after training, the machine was tested against the performance of actual prosecutors in the office, and its performance was equal to the average performance of the prosecutors. "In other words, this is a machine that is as good at being a prosecutor as an average prosecutor is."

The first question asked of participants was the trust item, which was analyzed and discussed in Study 6.4a. After the trust item, participants answered nine items of potential precipitating factors (see Table 6.1 or the Appendix for complete items). These nine items were presented in random order, except for one item that addressed a hypothetical and thus needed to be situated last. Participants were asked to what extent they agreed with each item, and the items were rated on a 6-point Likert scale that ranged from 0 (No, definitely disagree) to 5 (Yes, definitely agree).

One item discussed worries about data privacy and security. Another discussed general unease with having to deal with the adjudicator. Multiple items discussed perceived flaws in the process on account of: failing to account for situational nuance; failing to account for individual defendants' uniqueness; the system potentially having been designed by White people in order to perpetuate bias; suspicions that the adjudicator would discriminate; the adjudicator lacking necessary moral capacity; and inability by the adjudicator to adapt and improve over time. The

final item told participants to imagine that they had been arrested in Cooke County and asked if they believed that they would get a result from the adjudicator that would be better than what other defendants in their position would receive.

Table 6.1 Independent and Dependent Variables Used in Study 6.5

Indicator	Full item
Trust	I trust the performance of the [prosecutors / machine (the algorithm)] in Cooke County.
Privacy	I worry about data privacy and security given that defendants' information is processed by the [Cooke County Prosecutor's Office / machine (the algorithm) in Cooke County].
Unease	I think it's just weird and uncomfortable to have to deal with [a prosecutor / the machine rather than a real human prosecutor].
Situational nuance	I think the process is flawed: every situation is unique, and [prosecutors / the machine] can't possibly take into account all possible scenarios, such as intent, coercion, and so on.
Defendants' uniqueness	I think the process is flawed: every person is unique, and [prosecutors / the machine] can't possibly take into account individual differences, such as whether the defendant is a good person who just made a mistake.
Systemic racism	I think the [prosecutors / machine] will discriminate, and [they / it] will not be fair.
Expected discrimination	I think this sounds like [a system / an algorithm] designed by privileged White people in order to perpetuate, rather than diminish, racial bias.
Moral capacity	[Prosecutors have / The machine has] the moral capacity necessary for the job [they are / it is] performing.

Fear of lock-in	Thinking about the plea bargaining process in Cooke County, I am worried that the [prosecutors are / machine is] too inflexible and that this will create a barrier to change and future improvement in criminal justice.
Self-advantage	Now imagine that you have been arrested in Cooke County. To what extent do you agree with the following statement? I think I will get from the [prosecutor / machine] a result that's better than what other defendants in my position would receive.

Results

I included all of the measures for use in a multiple linear regression model. In addition to the above items, these included demographic features, and they also included adjudicator type as a binary variable. Thus, the baseline model included trust as the dependent variable, and the independent variables included all of the ones described above, which concerned data privacy, general unease, situational nuance, defendants' uniqueness, systemic racism, expected discrimination, moral capacity, fear of lock-in (inability of the adjudicators to adapt and improve over time), and self-advantage (belief that one will do better than other similarly-situated defendants). Also included as independent variables were adjudicator type (human or machine) and participant age, race, and gender.

To explore what was driving the dependent variable (the trust measure), I employed multiple linear regression estimation using ordinary least squares, and the model was determined via stepwise model selection by AIC (Bruce, Bruce, & Gedeck, 2017; James, Witten, Hastie, & Tibshirani, 2014). Results of the multiple linear regression yielded a model that well explained the variance in trust: $F(8, 183) = 17.4$, $p < .001$, $R^2 = .43$. There was a collective significant effect of the following features: situational nuance, expected discrimination, moral capacity, fear

of lock-in, and self-advantage (see Table 6.2). Adjudicator type and participant age also were significant.

Table 6.2 Multiple linear regression estimation using ordinary least squares, with stepwise model selection by AIC.

		Trust	
Predictors	Estimates	95% CI	*p*-value
(Intercept)	4.15	3.28 – 5.02	<0.001
Unease	-0.10	-0.23 – 0.03	0.131
Situational nuance	-0.21	-0.36 – -0.05	0.012*
Expected discrimination	-0.14	-0.28 – -0.01	0.039*
Moral capacity	0.23	0.10 – 0.35	<0.001**
Fear of lock-in	-0.18	-0.35 – 0.00	0.045*
Self-advantage	0.12	-0.00 – 0.24	0.056
Adjudicator type	0.39	0.05 – 0.73	0.024*
Participant age	-0.01	-0.03 – 0.00	0.009*

Note. *significant at $p < .05$. **significant at $p < .001$. Adjudicator type is a binary variable, where 0 = human and 1 = machine.

The effects accorded with the literature. As participants came to believe that the adjudicator was not taking into account the full nuance of the situation, their trust in the adjudicator decreased. As their belief that the adjudicator would discriminate increased, trust decreased. As their fear increased that the adjudicator would become locked-in and unable to improve over time, their trust in the adjudicator decreased. Finally, the more participants believed that the adjudicator possessed the requisite moral capacity for the legal decision making task, the more they trusted the adjudicator.

The main effects help shed light on these results (see Table 6.3 for all main effects). For those predictors that were significant, the machine generally scored worse. However, the only significant difference was for moral capacity: the machine was considered significantly less likely to possess the requisite moral capacity ($p < .001$).

Lastly, it is interesting that there were no significant differences in main effects by participant race (all $p_{adj} > .3$), except that, when the adjudicator was human, White (vs. Black) participants were more likely to believe that they would receive a better than average result from the adjudicator. See the Appendix (Appendix VI: "Table for Study 6.5") for the main effects broken down by participant race.

Table 6.3 Main Effects for the Independent Variable ("Trust") and the Dependent Variables

	Human		Machine		
	Mean	SD	Mean	SD	p_{adj}
Trust	2.70	1.07	2.50	1.35	>.99
Privacy	2.63	1.28	2.97	1.44	.82
Unease	2.50	1.30	3.58	1.21	<.001
Situational nuance	3.36	1.20	3.73	1.17	.34
Defendants' uniqueness	3.17	1.29	3.50	1.18	.63
Systemic racism	3.00	1.15	2.49	1.38	.06
Expected discrimination	2.82	1.56	2.56	1.54	>.99
Moral capacity	2.71	1.00	1.31	1.34	<.001
Fear of lock-in	3.09	1.22	3.56	1.12	.06
Self-advantage	2.54	1.20	2.03	1.16	.03

Note. Human = human adjudicator, that is, prosecutors. Machine = machine adjudicator. p_{adj} = two-tailed t-test with Bonferroni corrections.

Summary of Study 6.5

What tentative conclusions can be drawn from this analysis? In brief, it completes the picture. The model revealed that one particular feature relating to machine adjudication—moral capacity (Young & Monroe, 2019)—may shed light on the trust ratings. In particular, as participants' belief in an adjudicator's moral capacity increased, so did their trust in the adjudicator. As participants expressed greater fear that the adjudicator would be unable to change and improve its behavior, they expressed less trust in the adjudicator. Importantly, the machine adjudicator was rated much worse on moral capacity.

Discussion

I set out to ask whether machines are being held to a standard that is counterproductive: are lay individuals' preferences in legal adjudication leading them to disparage machines at the expense of innovations that can potentially reduce racial inequality in the justice system? The findings suggest an affirmative answer. In short, I showed that, at baseline, machines tend to be trusted less (Studies 6.1, 6.2, 6.3), but when it is made clear that they perform as well as humans, this lack of trust disappears (Study 6.4a). Said again, given equally unbiased performance, there was no difference in trust for human and machine adjudicators. However, as soon as errors appear, machines are disparaged (Studies 6.4b, 6.4c). Given equal rates of bias in human and machine legal adjudicators (that is, given equally suboptimal performance), participants had less trust in machine adjudicators. In other words, as soon as there was some bias, the machines were over-penalized for it. Even more, when human legal adjudicators showed twice as much racial bias as machine adjudicators, participants still trusted the machines less.

Furthermore, I found that this extreme discounting of machines that showed bias (vs. humans that showed the same or even greater bias) impacted support for interventions. Participants were much more likely to recommend shutting down the entire prosecutorial system until the bias was addressed when the bias was driven by machine, rather than human, adjudicators.

These results take steps towards explaining the conundrum outlined in the introduction to this chapter. Resistance to legal technology is significant and voluble, even though such tools show promise for decreasing bias, especially in comparison with bias that stems from human decision making. In short, it seems that, once any bias is identified within a machine legal adjudicator, the "first-failure effect" (Moray, 2003; Moray & Inagaki, 2000) and "algorithm

aversion" (Dietvorst, Simmons, & Massey, 2015) take over, and the machine is discounted beyond what is proportional or appropriate. That is, I found that human bias decreases one's trust in humans, but machine bias completely erodes one's trust in machines.

Beyond this effect—what I have been referring to as reverse legal automation bias—I found that there is one factor in particular that distinguishes trust in human and machine legal adjudicators. In particular, the model provided some evidence that moral capacity (Young & Monroe, 2019) might be particularly responsible for the trust ratings. In particular, as participants' belief in an adjudicator's moral capacity increased, so did their trust in the adjudicator; moreover, on this measure, the machine adjudicator was rated worse than the human adjudicator.

This is interesting because this feature is not a good guide for how much trust one should place in a legal adjudicator. Moral capacity may seem necessary for legal adjudication, but we must remember that courts have cited moral ratiocination as a reason for allowing wide human discretion, and human discretion has been shown to be a major driver of racial disparities. It may be the case that, for many kinds of legal decisions, clearer baselines, less deliberation, and less discretion will lead to outcomes that are more just.

Implications

Consider this statement from an AI research organization: "Using risk assessment tools to make fair decisions about human liberty would require solving deep ethical…challenges" (Partnership on AI, 2018, p. 3). This quote reflects an imbalance that courses through discussions of machine decision systems. It implies that for humans to make fair decisions about human liberty, no such deep ethical challenges must be solved—although we know that the very same

ethical challenges that attach to risk assessment tools also attach to the prosecutors, jurors, and judges who currently make these decisions (see Avery & Cooper, 2020b, 2020c).

In the context of legal adjudication, a primary difference between human and machine decisions that actually matters is one that has nothing to do with solving ethical quandaries; consider the difference in scrutability. Human motivations and aims are difficult to identify and even harder to address (Nisbett & Wilson, 1977). In contrast, machines are relatively more scrutable: computational purposes and processes can be declared prior to use and also verified after use (Kroll et al., 2017). In other words, one can know what a machine is doing and why it is doing it because one programmed it to do what it does. Even in cases of probabilistic or complex computation, such as when neural nets and other similar processes that are often called "black boxes" are used, there is an increasing number of techniques for unpacking how the machine made its decision (Lipton, 2016; Ribeiro, Singh, & Guestrin, 2016; Turek, 2019).

In short, we need to remember that ethical dilemmas are not unique to machine learning or to algorithmic systems. Human decision making demands the same tensions between prediction desiderata. In deciding whether to give someone a 5 or 10 year prison term, humans must decide whether to optimize the true or false positive rates, calibrate within groups (produce models in which the predictive parity of scores is the same across different demographic groups), and so on.[23]

Future Directions

There are many aspects that distinguish human from machine functioning, and one is intentionality. Humans sometimes act out of malice, profiteering, self-serving agendas. Indeed,

[23] Before turning to future directions, I must point out one limitation. These results might not be unique to the legal sector. While other sectors (we might think of finance or logistics) lack the moral element, and still others (we might think of healthcare) lack the extreme decisional uncertainty that inheres to legal decision making, I have not yet shown that this result in unique to the law.

bias may arise, in part, from such intentions. Machines do not act out of malice; they do not serve their own agendas. Frames that highlight this difference might lead to very different results: in essence, the obverse of the moral capacity result. Just as we want humans to make legal decisions because they are moral actors, we also might not want humans to make legal decisions because they are sometimes immoral actors.

In addition, people might prefer machine over human adjudication even when biased if they are reminded of the historical context. That is, if they are reminded that bias in the law is a human problem and one that has been with us since the U.S. Civil War, they might be more receptive not only to machines but to limited machine error as well.

In addition, there is a matter of construal that might be lurking behind these results. For example, if a self-driving vehicle crashes into a pedestrian, there is little to construe in the vehicle's intention, even if evasive maneuvers are undertaken. However, if a human driver crashes her vehicle into a pedestrian, we might construe evasive maneuvers as indicative of a desperate attempt to avoid harm, or a tortuous decision between multiple bad outcomes, or as an unfortunate byproduct of decision making under stressful conditions. Regardless, such construal might play a significant role in how the decision makers, especially in light of suboptimal outcomes like bias, are viewed. This is worthy of further exploration.

Lastly, in future work, it would be helpful to manipulate morality in an experimental design with the human/machine manipulation to confirm that it impacts trust. This would get us closer to a causal argument, and it would suggest ways in which machines might be "enlivened" in the public imagination by moral capacity attributions. In other words, if we can show that machines can perform tasks that approximate moral reasoning, the effects might extend beyond the specific context and influence overall trust.

Conclusion

Bias in the law is a durative and confounding problem. Interventions focused on correcting for human deficiencies have proven largely ineffective (Forscher et al., 2019), but technology has shown promise for decreasing bias in legal decision making (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018). Even so, there have been notable controversies around bias in legal technology, and it appears that the result has been to throw out the proverbial baby with the bathwater. The present research shows that not only is this overreaction occurring, it is the result of attitudinal backlash that follows machine failures. That is, while occurrences of bias in humans lead to decreased trust in the humans, occurrences of bias in machines completely erode trust in the machines. Moreover, I have shown that this backlash has potentially grave consequences in the struggle for racial equality: machines that are as adept at a legal decision making task and twice as racially fair as human decision makers are still trusted less than the humans. In addition, the humans are let off the hook and allowed to continue to operate with bias, even when they are twice as biased as the machines. This research suggests that, in order to achieve racial equity in the law, we may need to rethink our instinctive responses to machines and machine errors. Are we too hard on the machines? Perhaps so.

Note

The research presented in this chapter was conducted by Joel Cooper and the dissertation author.

Chapter 7

Decreasing Attorney Discretion through Algorithmic Decision Making—Examples from Criminal and Civil Law

In this chapter, I outline a method and process for how attorneys can use their own historical data to build computational models that enable them to anchor on racially fair machine generated outcomes. The driving insight is that, while we typically think of artificial intelligence in terms of automation (replacing human decision making), we instead should think of it as a second opinion, one that raises red flags and corrects human decisions that are straying towards bias. Using actual criminal case data, I provide a simulation of how this method would work in criminal law. Then, I discuss the development of an artificial legal intelligence that predicts the outcomes of civil legal cases, and I outline how it might be employed to decrease bias.

Introduction

Technology has become an integral component of myriad aspects of our lives. "Smart" technology locks household doors, answers telephones, drives cars and performs activities from the complex to the mundane. Algorithms determine whether baseball teams should use a shift in the infield as a particular batter faces off against a particular pitcher, and algorithms determine when our dishes are dry in the dishwasher. Technology can free people to focus on decisions that require judgment, and it can make decisions free of the foibles and biases that creep into human judgment. In this chapter, I will discuss the ways in which artificial intelligence, machine learning, and algorithms can aid decision making in the legal system. Such technology has the capability to help attorneys make decisions based on fair and unbiased use of data. By carefully considering how to properly use data-driven algorithms, the impact of conscious or unconscious biases in the law can be reduced.

A Cautionary Tale: The Mystery of the Biased Soap Dispenser

My enthusiasm for the use of technology to advance decision making in a fair and unbiased manner is not unlimited. It is tempered by an accompanying concern that the use of modern technology can result in some curious and functionally biased outcomes. In the summer of 2017, T.J. Fitzpatrick, who is Black, placed his hands under the soap machine at a Marriott hotel in Atlanta, Georgia. He received no soap. He assumed it was not working—until his White friend, Larry, used the same dispenser and found that it worked properly. Several additional attempts revealed that the dispenser was indeed discriminatory. It would work properly for White individuals but would not dispense soap for the hands of Black individuals.

Was the soap machine biased? No one contended that engineers intentionally used technology that favored White individuals over Black individuals, and no one believed that the Marriott intended to make it difficult for Black individuals to wash their hands. Nonetheless, Mr. Fitzpatrick could not wash his hands, and his White friend could. The soap machine was functionally biased. As it turned out, the dispenser was using near-infrared technology, which requires that hands presented to receive soap reflect light back to the sensor. Hands with darker skin pigments reflect less light, and so the dispenser failed to work for most Black individuals (Plenke, 2015). This is not bias in the sense of racial animus, but it does reflect an important fact: technology, in its various and multiple instantiations, is bound by rules. Those rules, inadvertently or not, can be applied equally to all and still produce unfair outcomes.

Research has shown that distrust of technology is increasingly widespread (e.g., see Metz, 2018). The AI Now Institute at New York University focuses much of its resources on identifying bias in artificial intelligence (AI) as well as when and how such bias impacts human rights and liberties (Whittaker et al., 2018). The Institute, which has partnered with the American

Civil Liberties Union, has expressed deep wariness of the increasing reliance on technology in modern societies (Whittaker et al., 2018).

At the same time, technologies that fall within the artificial intelligence umbrella, such as machine learning, and those that are much broader than what AI encompasses, such as predictive analytics, are increasingly in use. Organizations are expecting their managers to use data analytics to improve decision making. Ray Dalio, the billionaire hedge fund manager, has argued that organizations would benefit from recording almost all of their events, decisions, and outcomes as data (Dalio, 2017). Such a vision has been partially borne out in the medical realm: a team of researchers has taken more than 100 million data points from approximately 1.3 million medical cases and trained an AI model that performs nearly as well as experienced physicians at diagnosing diseases (Liang et al., 2019).

What does this mean for the law? We are nearing a key moment in the application of technology to legal processes. In the private sector, the legal technology market is huge: $3 billion (Mitratech, 2018) and poised to grow rapidly in the near future (Greentarget, 2018; Association of Corporate Counsel, 2019). In the public sector, there is an increasing push to use data in criminal justice (Mystal, 2019). But many decision makers, including prosecutors, are hesitant to implement technological advances. This hesitancy reflects the impact of activists and scholars who are arguing against the use of predictive analytics in criminal justice (Hao, 2019). Also, the hesitancy may reflect prosecutors' unease with computational decision making, as many prefer that decisions be left to attorneys. There is irony in such a preference, given that attorney experience and intuition have contributed to historical criminal justice disparities.

In this chapter, I outline the risk of racial bias in artificial intelligence. I then narrow my focus to bias in the application of technology to the law. In the succeeding two sections, I

provide a rejoinder, showing how technology has been used to correct for bias in the law, and I outline a framework for ensuring that technology is used properly. Lastly, I discuss two novel interventions that make use of technology and show great promise for reducing bias in attorney decision making.

Bias in Technology

Racist soap dispensers are notable curiosities. But there other, more significant instances of ostensibly racially biased technology—ones that have much greater impact. Google's facial recognition software infamously labeled two African Americans as gorillas (Zhang, 2015). There has been a string of recent incidents in which facial recognition software wrongly implicated Black individuals for crimes (Hill, 2020). Indeed, a number of facial-recognition systems have been shown to misidentify non-White faces at elevated rates (Garvie & Frankle, 2016). In order to understand what is going on in these cases, we need to understand the different ways in which bias in technology emerges.[24]

I begin with a prominent technology that powers much of the current AI use cases: machine learning. We can loosely define machine learning as when a computer learns to perform a task by being fed a training set of examples, after which it performs the same task with data it has not been fed previously (Louridas & Ebert, 2016). There are a number of different biases that affect machine learning. First, there might be sampling bias: when the population used to train the model does not accurately represent the population with which the model will interact. This is a major reason for the limited success of facial-recognition systems. Facial recognition models

[24] In this section, I am focusing on bias in technology, not on the lack of diversity in the technology sector, although the latter certainly may lead to the former. I leave that discussion for a more expansive treatment of the subject.

perform well at identifying White faces because they are primarily trained on White faces (see Buolamwini & Gebru, 2018).

Second, there is bias that emerges from base rates, which machines may interpret as reflective of causal relationships. For instance, doctors may statistically be more likely to be male, but a machine that always concludes that doctors are male (or that a female could not be a doctor) would be making a significant error. Something along these lines occurred when Amazon attempted to automate its resume-screening procedures (Dastin, 2018). Because many of Amazon's early, successful employees had been male (i.e., the base rate was heavily male), the machine screened out nearly all female candidates. In short, the machine's bias was the result of it picking up on the base rate that humans had created.

Indeed, different base rates across different legally protected groups is a major complication for fairness in predictive analytics (see Berk et al., 2018). It is a primary factor in the so-called "group fairness impossibility theorem" (see Chouldechova, 2017), which arises from the fact that different groups nearly always have different base rates, and different base rates, absent perfect prediction, will lead to different error rates (we can think of false positives and false negatives). Pursuing this line of thought, Kleinberg, Mullainathan, and Raghavan (2017) outlined various mathematical definitions of fairness that could reasonably be used in the criminal justice context—and they showed that these three definitions could not all be operationally satisfied at the same time. Likewise, Verma and Rubin (2018) discussed twenty different definitions of fairness.

Even though "the detailed differences between [these] multiple definitions are difficult to grasp" (Verma & Rubin, 2018, p. 1) and full explication is beyond the scope of this chapter, it is illustrative to go over a few examples. Let's consider the case of determining a defendant's risk

score. For simplicity, let's assume that the score is either a "good" risk score (where good means that the defendant is unlikely to commit additional crimes and likely to show up to future court dates) or a "bad" risk score (likely to commit additional crimes and unlikely to show up to court). A true positive is when a defendant is given a good risk score, and this score accurately reflects the reality (the defendant is unlikely to commit additional crimes and likely to show up to future court dates). A false positive means that a defendant is given a good risk score, but this is not accurate (i.e., the defendant should have been given a bad risk score). A true negative is when a defendant is given a bad risk score, and this accurately reflects the reality. A false negative is when a defendant is given a bad risk score, but this does not reflect the reality (i.e., the defendant should have been given a good risk score). In any prediction task, we can group outcomes into a positive predicted class, which includes the true positives and the false positives, and we can group outcomes into a negative predicted class, which includes the true negatives and the false negatives.

One definition of fairness is "group fairness." We can say that a classifier satisfies this definition if both protected (e.g., Black defendants) and unprotected (White defendants) groups have equal probabilities of being assigned to the positive predicted class (Verma & Rubin, 2018). This is intuitively fair. Since positive prediction means a good risk score, then all defendants, regardless of group, should have an equivalent opportunity of receiving a good score.

But group fairness considers only predicted outcomes; it fails to include a consideration of actual outcomes. A second definition of fairness, "predictive parity," considers both predicted and actual outcomes. A classifier would satisfy this definition if both Black and White defendants have equal positive predictive value (PPV); that is, for both Black and White defendants, the probability that a defendant who received a good score will actually refrain from

additional crimes and show up for subsequent court dates should be the same. This is an intuitively appealing definition of fairness: the percentage of *correct* positive predictions should be the same across race/ethnicity, or else it would seem that one group was receiving preferential treatment. While this definition is appealing, the problem is that it may be in conflict with the group fairness definition. It is possible to have predictive parity but fail to satisfy group fairness—and vice versa. And these are just two of the fairness definitions; there are many more equally intuitive definitions of fairness, and they also might conflict with each other (see Verma & Rubin, 2018).

While research in algorithmic fairness is moving quickly and meeting with some success—researchers like Noriega-Campero and colleagues (2019) are beginning to find ways to simultaneously satisfy some of the more common definitions of fairness—it is true that definitions of fairness will often be in conflict. This merely draws into relief the fact that the conflict is a philosophical, political, or, most truly, a policy one, and it requires that policymakers must work with data scientists to decide what is best.

In addition to fairness concerns, there also are concerns about whether the results of machine decisions are appealable. Machine decision making has been used in disability determinations, where there is a risk that outcomes will unduly burden some groups and that those groups will have no recourse. In a recent case, individuals who were adversely affected by machine decisions were not given insight into the machine's decision making and had no means of recourse, such as an appeal (Lecher, 2018). Whether or not this appeal/remedy problem will be addressed legislatively is still unclear (Bowles, 2019).

In the next section, I discuss bias in legal technology. As prelude to that discussion, it is worth focusing for a moment on one instance in which technology within criminal justice is

allegedly biased: predictive policing, a concept that captured the public imagination with the release of the Steven Spielberg film, *Minority Report*. A few years ago, the Los Angeles Police Department implemented a new program to combat property crime (Rieland, 2018). The program involves reliance on a software system—PredPol—that instantiates an algorithm that examined 10 years-worth of crime data, including the types of crimes committed, dates and times, and locations. The software produces a prediction as to where and when crimes are likely to occur over the succeeding 12 hours. In essence, PredPol creates "hotspots"—areas of likely criminal behavior—on which police can focus their resources. It has been alleged by critics that the software provides a patina of objectivity to what is mere bias: officers want to over-police predominantly Black neighborhoods, and the software gives them an explicit reason to do so (LA Times Editorial Board, 2019).

Bias in Legal Technology

Machine learning has been used for legal prediction for more than 50 years (e.g., see Mackaay & Robillard, 1974). While most of the uses have been academic or have involved private sector applications, there is one major area of concern in public sector use: statistical algorithms that have been employed in risk assessment. For instance, in New Jersey, there is a pretrial assessment system designed to algorithmically generate "risk" scores that indicate whether a defendant should go free or be detained in jail while awaiting trial (Hanna, 2018; O'Dea, 2018). While there are many different risk assessment systems, all of them use rules and/or machines to generate outcomes (recidivism scores; dangerousness scores) that impact individuals' freedom and well-being.

Such algorithms have been charged with perpetuating racial bias (Angwin, Larson, Mattu, & Kirchner, 2016; Barocas & Selbst, 2016). Starr (2014) wrote that these algorithms may

work to hide "overt discrimination based on demographics and socioeconomic status" (p. 806). A 2016 report by *ProPublica* alleging that an algorithm used in Florida—called the COMPAS system—was biased against Black defendants received nationwide attention (Angwin et al., 2016). *ProPublica* analyzed 10,000 cases in one Florida county and found that Black individuals (compared with White individuals) were almost twice as likely to be scored as high-risk to reoffend but then not reoffend. White individuals seemed to get off easy. They were more likely than Black individuals to be labeled low-risk but then reoffend.

Much of the controversy around the COMPAS system is a matter of statistical tests and is more complicated than suggested in the previous paragraph. When it comes to predicting who will reoffend, a true positive is when someone who is scored as likely to reoffend does reoffend. A false positive is when someone who is scored as likely to reoffend does not reoffend. A false negative is when someone who is scored as unlikely to reoffend does, in fact, reoffend. In analyzing the performance of the COMPAS system, *ProPublica* analyzed false positive rates and false negative rates, finding that White individuals were treated better. COMPAS's parent company analyzed positive predictive value, which places more weight on true positives, and the results showed that Black and White individuals were treated equally. Which test is superior is a matter of perspective (see Spielkamp, 2017).

Issues involving the use of COMPAS even made it to the Wisconsin Supreme Court (*State v. Loomis*, Wis. 2016), where it was held that judges can use computer aided risk scores, but they cannot comprise judges' sole consideration. In spite of this seemingly permissive stance, some have argued that the Court tried to temper enthusiasm for automated risk assessments (see Note, 2017). In a concurring opinion, the Chief Justice wrote that while "consideration of

COMPAS is permissible; reliance on COMPAS for the sentence imposed is not permissible" (*State v. Loomis*, Wis. 2016, p. 774).

Legal Technology for Good

In spite of technology's potential for bias, I nonetheless believe that the proper use of technology can help correct for the human biases that have plagued the criminal justice system. First, racial bias can be either explicit or implicit. Broadly, we might say that explicit racial bias refers to consciously held racial attitudes, while implicit racial bias refers to racial attitudes that are less readily available to introspection (Greenwald & Banaji, 1995) and less subject to conscious control (Greenwald, McGhee, & Schwartz, 1998). Given that explicit racial bias triggers equal protection concerns, individuals working in public settings are relatively unlikely to reveal explicit racial bias. Thus, most interventions focus on what is less conscious and less visible: implicit bias. But implicit bias interventions have shown limited efficacy. A recent meta-analysis found that, while some changes in implicit bias can be achieved through interventions (at least in the short term), most interventions do not work. Of those that do, the change in implicit bias is generally small, and any subsequent change in behavior is negligible (Forscher et al., 2019).

Interventions besides those focusing on attitudes have also met with limited success. Smith and Levinson (2012) recommended hiring a more diverse pool of assistant district attorneys, but studies show that in-group bias might be just as great, if not greater, than out-group bias (e.g., see Depew, Eren, & Mocan, 2017; Eberhardt, 2019). In her 2019 book, Eberhardt tells a story in which her son, who is African American, observed a Black man boarding an airplane and opined that the man might be dangerous or a criminal. Depew and colleagues (2017) provided the first empirical evidence of negative in-group bias in a

randomized design outside the laboratory. They reviewed Louisiana juvenile court cases from 1996 through 2012 and found that, all else held equal, Black (vs. White) juveniles who were randomly assigned to Black (vs. White) judges received longer sentences and were more likely to be incarcerated rather than given probation.

Another commonly cited solution is to blind criminal justice actors to defendants' race (see Sah, Robertson, & Baughman, 2015). This practice is currently being implemented by the San Francisco District Attorney's Office (Gecker, 2019). But there are limitations to such an approach: even if a defendant's skin color is hidden, cues and correlated variables, such as defendants' zip codes, names, or colleges attended, may still be present (Sah et al., 2016). In addition, prosecutors must see defendants at some point, so blinding can only impact initial case decision making.

Given that there is significant bias in U.S. criminal justice and almost no interventions have successfully curtailed it, the time is right to introduce technology as a way to correct for human error. For some years now, it has been known that big data and artificial intelligence often outperform human judgment alone (Berk, 2008; Lazer et al., 2009). Improvements in decision making have been shown to be possible in a number of domains, including workplace safety (Kleinberg et al., 2015; Morantz, 2008), taxation (The Department of the Treasury, 2009), municipal government (Simon, 2014)—and even criminal justice (Berk, 2008).

There are a number of examples of existing technological applications that are having positive impacts on legal processes and outcomes, including on lessening bias. Recently, two California counties teamed up with Code for America to automate the clearance of more than 50,000 marijuana convictions that had been made eligible for clearance under a newly enacted proposition (Ready, 2019). An algorithm combed through government files to identify eligible

individuals and file the motions to clear. The automation was thorough: the person with the criminal record did not have to file a petition, hire an attorney, or attend a hearing—everything was completed by the machine (Ready, 2019). Because African Americans were overrepresented in this population, one can say that the technology was working to correct for historical bias.

This is not the only example of legal technology working for good. As discussed above, in San Francisco, AI is being used to blind police reports so that prosecutors cannot tell the race of defendants (Gecker, 2019). Jurisdictions across the country are using the Public Safety Assessment tool (PSA), which was designed to reduce bias in judicial decision making. In assessing dangerousness, the tool takes into account age and history of criminal convictions, but it elides race, gender, employment background, where a defendant lives, and history of criminal arrests (Arnold Foundation, 2018). Finally, there is groundbreaking theoretical work showing that machine predictions can lead to drastically improved legal outcomes (Kleinberg et al., 2018). The authors found that their policy simulation would reduce crime by up to 25% without causing an increase in jail populations. Alternatively, it would be possible to keep crime rates steady while reducing jail populations by 42%.

Essentials for Using Legal Technology to Correct for Bias

So far in this chapter, we have seen how technology, including legal technology, can lead to poor outcomes, such as racial inequality. We also have seen how legal technology can be used to correct for racial bias. In this section, we embrace this latter possibility—legal technology for good—and outline five essential tasks that should guide any such legal technology project.

I. Improving Data Collection

First and foremost is to improve data collection. No system that relies on data can lead to reliable outcomes if the data are not complete and accurate. When it comes to the law, and

especially criminal law, data collection is notoriously poor (Executive Office of the President, 2016). Many offices, such as prosecutors' offices, keep few data, and what data they do keep, they try to keep confidential. This has been the norm for many years. That said, change is afoot. Connecticut just passed legislation requiring prosecutors to compile data on how many defendants received prison time, plea bargains, or were diverted, and the data must be parsed by race, ethnicity, sex, and age (Mystal, 2019). Other prosecutors' offices around the country are making similar efforts to collect data that is necessary for determining whether similarly-situated defendants are being treated similarly (Kutateladze et al., 2018; Pantazi, 2018).

But data collection alone is not enough. The quality of the data collected is important. A 2018 report by the Urban Institute found that while prosecutors are increasingly using data to inform decision making, concerns about data accuracy may be standing in the way of widespread adoption of such use (Olsen, Courtney, Warnberg, & Samuels, 2018). This is a serious concern, one that has been echoed by critics of data science in the law (e.g., see Hao, 2019).

I emphasize that, just as the Connecticut legislature recognized, data must be collected by prosecutors' offices, but the collection does not have to be exhaustive. It simply must fit the stated goal: data must be consolidated and aligned for the purpose of addressing racial bias. Only those variables that are essential for assessing bias are needed, including but not limited to initial charges, case outcomes and method of resolution, as well as demographic information on defendants, prosecutors, judges, and arresting officers. Regardless of whether an office has historically been deficient at such data collection, proper collection should begin moving forward.

II. Expanding Data Analysis Capabilities

The second essential task is to expand data analysis capabilities. Prior to beginning this discussion, it is worth noting that, even though some offices, such as the District Attorney of New York, are equipped with a data analysis unit, many prosecutors' offices are not so equipped (Varghese, 2019). Performing the requisite analyses, building models, and instantiating those models in software are big tasks, and I outline how they might be achieved in the discussion below on implementation. Here, I focus solely on the specifics of what such analysis should include.

If an office has completed the first task (data collection), it then can turn to analysis. Analysis can be in service of any of a number of different objectives. An office may want to analyze whether its prosecutors are treating similarly situated defendants similarly. An office may want to use predictive analytics to help guide prosecutorial decision making (for more on this use case, see the final section of this chapter). An office may want to use analytics to identify weaker from stronger cases, thus creating efficiencies in determining when to dismiss and when to pursue a matter further. In short, there are numerous reasons for analyzing properly collected criminal justice data.

In our previous discussion of the COMPAS system, we considered how training a model on data that reflects base rates can lead to incorrect conclusions about causality. An example of this is the statistical fact that most nurses are female. To correct for bias of this type, it is necessary to involve experts who are sensitive to the social context in which the AI will be operating (West, Whittaker, & Crawford, 2019). If the domain is law, and the machine is evaluating drug offenses, then lawyers who understand historical racial bias and the fact that minorities—and especially African Americans—have been over-policed and over-prosecuted for marijuana offenses, must be involved in critiquing the models' outputs.

No matter how well-collected the criminal justice data, any technology that builds upon it must be vetted by experts who can contextualize the data. This is necessary because, at the least, technology should operate fairly, and fairness must be defined by key stakeholders. Once it is so defined, then context becomes vital for making sure that the legal technology employs the fairness definition that the key stakeholders think is the appropriate one to use (see Selbst et al., 2019). For example, how criminal records should be weighed in analyses of future criminal risk is a policy question. Once it is answered by policy experts, then the instantiation will require careful reworking of the algorithm.

When the legal technology employed makes use of predictive analytics, there are numerous methods for addressing bias in the historical record (see Corbett-Davis et al., 2017; Hardt, Price, & Srebro, 2016; Kamiran, Zliobaite, & Calders, 2013). Many of these methods rely on determining which predictors are legitimate (e.g., which ones are valid in determining predicted recidivism) and which are illegitimate (e.g., race or other features—such as zip code—that might function as mere proxies for race). While we just gave a few examples, doing so elided that it is difficult to determine which predictors are legitimate. Harcourt (2008), for instance, argued that criminal records are proxies for race. Is this true? An argument can be made that criminal convictions are legitimate predictors but criminal arrests are not, since the latter might reflect policing bias. With such an approach, in subsequent generations those with clean records will be treated equally; but this approach acknowledges that not all past injustice can be corrected with one algorithm. It should be clear that decisions like this one—how to handle criminal records—require deep knowledge of the area of expertise (Ho, 2018).

Pope & Snydor (2011) have shown the effectiveness of using information from illegitimate predictors but debiasing the predictors by marginalization. To flesh this out, their

method involves first estimating a full model that includes any and all predictors (crime charged, race, criminal history, precinct of arrest, and so on). Then they identify variables that are illegitimate: those that are protected (such as race) and also those that are potentially problematic (such as zip code, which may correlate with race). When the model is subsequently used to produce a prediction for an individual, the coefficients for the protected and potentially problematic variables are not used. Instead, they are replaced with their population means. For example, if a model is estimated with a dummy variable for minority status (0 = non-minority and 1 = minority), then at the predictive stage the minority variable will be set to the population mean of minorities, rather than 0 or 1 (Pope & Snydor, 2011). The strength of this method is that prediction is maximized because all useful variables are used in creating the estimations, but protected and potentially problematic variables do not impact outcomes for individuals. Similar methods have been proposed and refined by Mancuhan and Clifton (2014) and Lipton, Chouldechova, and McAuley (2017). All of these approaches build from the recognition that we should actively identify and correct for racial or other improper disparities: not only does blindness not necessarily entail fairness (Dwork et al., 2012), it often is a poor notion of fairness (Hardt, Price, & Srebro, 2016).

"Proportional bias" also should be taken into account. For example, there may be an underrepresentation of White drug offenders in a given data set. To deal with such disproportionate data, Amini and colleagues (2019) created an algorithm that combines an original learning task with a variational autoencoder in order to learn the latent structure of the dataset. The algorithm then uses the learned latent distributions to reweight certain features during training—a process that can be used to address proportional racial and gender bias (see also Chen, Johansson, & Sontag, 2018).

There are reasons to be cautious about employing technical solutions to debias algorithms (Campolo et al., 2017), and there definitely should be direct challenges to techniques developed for debiasing algorithms (see Altenburger & Ho, 2018), but these are reasons for being cautious and for acknowledging hurdles. They are not sufficient to forgo the use of legal technology altogether.

III. The Importance of Implementation

The third essential task for using legal technology for good is to pursue proper implementation. By "implementation," I mean the specific factors that determine how legal technology is brought into daily office practice. Leadership is the first area to address. Whenever the instituting body is governmental, we can assume that there will be up to three primary groups working on the project: the office's leadership (e.g., an elected/appointed district attorney and his or her senior staff); the office's employees (e.g. assistant district attorneys and legal assistants); and a data analytics team, which will be an in-house team, an academic team working from outside the office, or (most likely) a combination of in-house data analytics people and outside academics.

A number of important objectives will fall to the leadership team. First, there will be a need for AI standards, processes, and policies. This includes a need to address data privacy and protection concerns, of course, but it also involves addressing theoretical queries. Above, I discussed that there are many definitions of fairness. I also discussed that, within fairness discussions, there are many questions (such as which aspects of criminal histories to include) that require policy answers. While these thorny questions may be flagged by the data analytics team, it is the leadership team that likely will have to answer them.

The leadership team also will be responsible for providing a clear vision for the office. A recent report found that, while pretrial risk assessments may lead to biased outcomes if used carelessly, targeted and teleological use leads to both significantly reduced pretrial detention *and* less racial disparities—in short, a win by all accounts (Picard, Watkins, Rempel, & Kerodal, 2019). "Teleological" means that the use must be guided by clear goals, such as decreased disparities, increased efficiency, or fewer incarcerations. "Targeted" means that leadership must consider whether the tool should be used only in a subset of cases, such as those that are most closely related to the stated goals (e.g., if the goal is to decrease incarceration, then there should be a focus on violent crimes committed by high-risk offenders).

For our topic of interest (racial bias), the vision is that AI is going to be used to decrease racial bias in the office (there also may be secondary and incidental goals, such as rendering decision making more efficient and consistent). This vision—an office without bias—must be presented to the office in an engaging and honest manner, one that captures the fact that racial bias poses a deep and almost existential threat. Bias in criminal justice against minorities and darker-skinned individuals was significant historically, and it is significant today. AI shows the best potential for solving this problem—this is an example of one possible vision; note that it is generated by the leadership team and seeks to unite both leadership and employees on a common and exciting goal.

Further, the leadership team likely should emphasize that legal technology will not be used to replace attorneys, nor will it be used to replace attorney judgment. As Fountaine, McCarthy, & Saleh (2019) have written for the private sector, "When AI is adopted broadly, employees up and down the hierarchy will augment their own judgment and intuition with algorithms' recommendations to arrive at better answers than either humans or machines could

reach on their own" (p. 65). Indeed, voluminous research has found that professional judgment improves when it is coupled with data-based analyses and guidance (Dawes, Faust, & Meehl, 1989; Gendreau, Little, & Goggin, 1996; Kuncel, Klieger, Connelly, & Ones, 2013). This combination of human decision making informed by cognitive computing is sometimes called "collaborative intelligence" (Wilson & Daugherty, 2018), and the term rightly captures how legal technology should be used: to empower attorneys by placing them in a working partnership with AI. As the CEO of a multi-billion-dollar company wrote in the *Harvard Business Review*, "A good person plus a good algorithm is far superior to the best person or the best algorithm alone" (Lake, 2018). For example, researchers employed a deep-learning algorithm to detect metastatic breast cancer from slides of biopsies, and the algorithm was accurate 92.5% of the time (Kleinberg et al., 2018). A physician who reviewed the same slides was accurate 96.6% of the time. However, when the two systems were combined—algorithm and physician—the accuracy rate rose to a near-perfect 99.5%.

Lastly, much research shows that leadership should pay at least as much attention to adoption of the technology as they do to the technology itself (Fountaine, McCarthy, & Saleh, 2019). This is especially true when it comes to legal technology, since the legal profession has been slow to adopt new technology (Cohen, 2018; Donahue, 2018), an unsurprising fact given that, in most business that are not "born digital," successful AI implementation is often an uphill battle (Fountaine, McCarthy, & Saleh, 2019, p. 65). To encourage adoption, leadership should provide incentives for employees who facilitate deployment of the legal technology undertaking. Leadership also should shine a spotlight on top performers (Fountaine, McCarthy, & Saleh, 2019); for example, prosecutors who embrace the new legal technology and volunteer to pilot it

should be recognized publicly and also considered for new roles, perhaps ones that span the employee and data analytics teams.

This brings me to the need for translators. Either in-house individuals or collaborating academic researchers could bridge the gap between the data engineers and the employees (often attorneys). These translators can communicate the purpose of the underlying technology. This is distinct from communicating the vision of the project, which falls within leadership's remit. Translators, by contrast, can explain the intricate details of the technology and why it is producing the output that it is producing. In short, translators can explain the technology to employees and inform data analytics team members of employees' complaints and suggestions when they arise.

Translators also will play a key role during implementation, studying the workflow to diagnose and troubleshoot problems that arise. Fully integrating transformative legal technology into an office's processes is a long journey. Machine learning and other techniques that comprise legal technology are more art than science. Successful legal technology requires experimentation, and it is certainly the case that early models will not perform optimally. It also is the case that, once a model is working well, attorneys will struggle to use it properly. In all such instances, translators must highlight what was learned from each prior iteration and present the findings to leadership.

IV. Transparency

The fourth essential task for using legal technology to decrease bias is to be transparent. Offices that are using legal technology have a duty to inform the public of the use. They should publicize when and where AI systems are used—and for what purpose. More importantly, almost all legal technology makes use of computer code. Some of this code, and some of the algorithms,

are exceedingly complex and difficult to parse rationally. For instance, neural nets are powerful tools, but they are not programmed to produce consistent or replicable results. In short, it can be hard to understand the machine's decision process. This raises deep questions relating to the imperative—long established in U.S. law—that governmental decision making be explainable (See Crawford & Schultz, 2014; Selbst & Powles, 2017). Thus, offices should hesitate to use fully opaque technologies (See Citron & Pasquale, 2014; Edwards & Veale, 2017); and, if any such technology is used, it should be analyzed by researchers working on model explanations, and the analyses should be relayed to interested or affected parties. Explaining complex computational models may seem like a Sisyphean task. In *State v. Loomis* (2016), the Wisconsin Supreme Court case discussed above, Justice Abrahamson wrote in a concurrence, "[T]his court's lack of understanding of COMPAS was a significant problem… The court needed all the help it could get" (p. 774). But this is a failure to explain, not a problem of explainability. Lakkaraju and colleagues (2019) have shown that, even with extremely complex models, there are methods of dividing the models into smaller subspaces which can be explained and, in turn, allow for intelligible explanation of overall model workings.

V. Accountability

The final essential task is to be accountable. This consists of setting clear metrics for success, duratively assessing progress on those metrics, allowing for appeals, and making data and code available for outside review and auditing. The first two prongs—setting clear metrics and assessing success by them—are intimate with the goals set by leadership. If the goal is to rid the office of racial bias in case outcomes, then the question undergirding the metrics is clear: are similarly-situated defendants being treated similarly? While there are a few different ways of assessing this, the office should decide which ones to take (and using more than one is certainly

appropriate), and it should produce quarterly reports, similar to a financial accounting, that describes how the office is doing in terms of these metrics.

Some researchers, such as Reisman, Schultz, Crawford, and Whittaker (2018), have argued for Algorithmic Impact Assessments (AIAs). AIAs would require extensive review of technological impact, and this review would be independent of the office's stated goals. So, if the goal of the legal technology is efficiency and not racial fairness, fairness concerns would still be monitored pursuant to AIAs (see also Selbst, 2017).

While offices seldom like appeals or disturbances of past decisions, there should be a mechanism in place for those impacted by legal technology to contest the results. This does not have to be an expansive right; it should be tailored to accord with the right an office grants for appealing human decisions.

Lastly, accountability means that offices employing legal technology should make both data and code publicly available. The data will be partially scrubbed and redacted for privacy reasons, but as much as possible should be disclosed. In this way, auditing and outside review will be a matter of will only. This is not a mere utopian vision, either. Some offices have already begun to comply with such disclosure principles: the district attorneys' offices in San Francisco and Chicago have posted much of their data to public repositories so researchers can see precisely how defendants are being treated (see DA Stat, 2019; Cook County Government, 2019)

Rethinking the Aim of Artificial Legal Intelligence

While we typically think of artificial legal intelligence in terms of automation (replacing human labor), we instead might think of it as a second opinion,[25] a guide that can raise red flags

[25] Indeed, the healthcare sector is far ahead of the legal sector in incorporating AI as a second opinion (Chi-Tung et al., 2020; Lan, Zhou, Duan, & Yan, 2018).

and correct human decisions that stray towards bias. In other words, if a system is designed in accordance with the five principles outlined above, it then might be implemented as a guardrail of sorts, one that prevents deviations into bias without eliminating human judgment and decision making altogether. What would this look like? In the next two sections, I present examples in criminal and civil law.

A Potential Use Case of Technology to Correct for Bias in Criminal Law

In this chapter, I have discussed how bias in technology can be a significant problem and how the same is seemingly true of bias in legal technology. However, given the inexorability of human biases, the long history of disparate treatment of racial minorities, and the promise of current legal technology, it makes sense to turn to legal technology to correct for bias in the law. In the previous section, I outlined essential tasks that, if accomplished, would set the stage for using legal technology to accomplish this purpose.

Some might argue that, even if the steps outlined above are followed, there are insurmountable problems with using legal technology to decrease bias. For instance, we know that when there are similarly situated male and female defendants, male defendants receive more severe sentences. If a prosecutor's office built out their data collection and analysis capabilities, they would be able to quantify the extent of this pro-female bias. But if they installed an algorithm that corrected for this bias, this could be considered a violation of the Equal Protection Clause (see Starr, 2014; Oleson, 2011). If we imagine parallel and analogous algorithmic uses, we might find violations of Title VII. Indeed, it seems that algorithmic correction would be a form of discrimination—or so goes the argument.

I have three rejoinders. First, if there are disparities, these need to be documented. The fact that legal technology might not be able to correct the disparities does not mean that it should

not be used to document them. Second, if the algorithm described above violates some principle of non-discrimination, then it must be the case that the attorneys who caused the disparities in the historical record also violated a principle of non-discrimination. Given this, it does not make sense to ban the former, which is transparent in its quest for non-discrimination, and allow the latter, which is opaque in its goals: after all, prosecutors are likely not aware that they are biased in their decision making, even if the outcomes suggest that they are. Third, there are countless ways to employ legal technology that would not run afoul of non-discrimination principles. The algorithm could simply treat everyone like the best-off group: that is, male defendants would be treated like female defendants, and female defendants would be treated like female defendants. Alternatively, an average across the groups could be established and everyone could be treated according to this average: male defendants would be treated like male/female average defendants, and female defendants would be treated like male/female defendants. As another alternative (and this is the one that I recommend), the model could be used in a strictly advisory capacity. It is a second opinion, providing guidance, and prosecutors would be told that it had been reweighted to correct for previously documented bias.[26] Regardless of which path is taken, it is certain that moving away from data and away from legal technology is not a solution. I think it would be a mistake, as it ensures that the status quo, where disparate outcomes caused by biased humans is the norm, will remain.

To help clarify what a legal technology intervention might look like, I now provide something of a coda: an example of two ways in which legal technology could be used to correct for bias. For the first way, I describe a project to develop a predictive model that can render

[26] This solution nods in the direction of a related issue: that the European Union's General Data Protection Regulation (GDPR) specifies that individuals have the right not to be subject to decisions based *solely* on automated processing (European Commission, 2018). See also *State v. Loomis*, 881 N.W.2d 749 (Wis. 2016), which permits judges to consider automated decision making so long as consideration does not rise to full reliance.

prosecutorial decision making more equitable and resistant to racial bias. The project involves building a model that, based on initial case information that a prosecutor receives, suggests how the case would resolve if the defendant were treated race-neutrally.

A "suggestive model" to aid prosecutorial decision making. As tempting as it is to think that each criminal case is unique and that prosecutorial work is a nuanced undertaking, prosecutors often operate with limited information about what actually transpired, and this may lead to reliance on extralegal influences, such as racial stereotypes (Kutateladze, Lawson, & Andiloro, 2015). An analysis of 2010 and 2011 felony drug cases in New York City found that, of numerous evidentiary factors considered, almost none mattered for predicting sentence offers (Kutateladze, Lawson, & Andiloro, 2015). In addition to limited information, prosecutors also operate under time constraints, another factor that has been shown to lead to reliance on extralegal factors (Albonetti, 1991; see also Kahneman, 2011, for a discussion of reliance on heuristics). Third, and most importantly, a difficulty inherent to legal work generally and prosecutorial work specifically is that decisions often must be made in the absence of clear baselines. This is especially true when a prosecutor is determining which charges best attach to a specific fact pattern. Given these constraints, prosecutors are especially liable to overreliance on stereotypes, such as the association of African Americans with violence and criminality (e.g., see Hawkins, 1981; Steffensmeier, Ulmer, & Kramer, 1998).

Prosecutors are less likely to offer Black individuals a plea bargain, less likely to reduce charge offers for Black individuals, and more likely to offer Black individuals plea bargains that include prison time (Albonetti, 1990, 1992; Frenzel & Ball, 2007; Kutateladze, Andiloro, & Johnson, 2016; Kutateladze, Andiloro, Johnson, & Spohn, 2014; McDonald & Raphael, 2017). Defendants who are Black, young, and male fare especially poorly (Albonetti, 1990; Kellough &

Wortley, 2002; Kutateladze, Andiloro, & Johnson, 2016). In a comprehensive study of federal cases, Rehavi and Starr (2014) found that Black individuals received sentences that were 10 percent longer than those of similarly-situated individuals and that this disparity was largely explained by prosecutors' charging decisions.

Fortunately, prosecutors' offices are in possession of information that can be used to form clear and unbiased baselines: closed casefiles. Employing advanced methods of statistics and computer science, they could use closed casefiles as a corpus from which to build a model that can produce race-neutral case outcomes. Specifically, they could build a computational model that enables introduction, early in the development of a new criminal case, of an anchor to prosecutorial decision making. That is, the model would receive the police report and would produce a prediction as to how the case would resolve (based on tens of thousands of previous cases) if the defendant were treated race-neutrally (that is, all else held equal, the defendant will not receive a different outcome if he is, say, White rather than Black). This prediction could be used by the prosecutor as a second opinion, as an anchor on more fair and just outcomes.

The first phase of this endeavor would involve data collection and initial analysis. The goal of Phase I data analysis would be to determine whether and to what extent the office's prosecutors have exhibited racial bias.

In the second phase, the office will build a model that, based on arrest charges, the arresting officer's narrative report, and the defendant's criminal history, can produce a prediction as to how a case would resolve if the defendant were treated race-neutrally.

Machine learning involves using feature frequency information to statistically map correspondence between features and outcomes. For the present purposes, these are case features (arrest charges, arrest report, etc.) and case outcomes (in most instances, the plea bargain

reached). Of course, case features also include defendant characteristics, such as age, gender, and race. Thus, if one seeks to build an optimally predictive model, race should be included in the model (see Mayson, 2019; see also Bartlett, Morse, Stanton, & Wallace, 2018). Indeed, this approach, in which bias is a feature, is the usual one when it comes to predictive analytics in the law.

The approach I am suggesting here is distinct, as the goal is to remove suspect features, such as racial bias, from the model, thus bringing its predictions into closer accord with Constitutional mandates for racially equitable treatment of criminal defendants by state actors. Rather than predict likely outcomes, the objective for the model is to suggest how the case would resolve if the defendant were treated race-neutrally. For this reason, we call the model a "suggestive model."

Scholarly precedents for this work include those reviewed in the data analysis discussion above. Moreover, the key to such model construction is ensuring that similarly-situated defendants truly are treated similarly, keeping in mind that, while machine learning algorithms lack conscious intention to be biased, proxies for race may still lead to models that reflect historical patterns of discrimination (Weisberg, 2018; see also Dressel & Farid, 2018). Thus an important and exciting part of the work would be the handling of covariates. This was true in research we conducted with data from a large, urban prosecutor's office (see below), wherein I employed a similar approach in building a model that predicts charge reductions (i.e., when a defendant who is initially charged with a felony winds up pleading guilty to a lesser charge). Not all defendants are treated as if they are White, but all similarly-situated defendants are treated similarly (see Aliverti, Lum, Johndrow, & Dunson, 2018; Johndrow & Lum, 2018). With such a model, Black defendants on average receive more lenient treatment than a pure Black-trained

model would predict, and White defendants on average receive more severe treatment than a pure White-trained model would predict, bringing into relief the above-mentioned fact that the power to be lenient is the power to discriminate.

The third phase of the research program would involve working with the prosecutors to ensure that they are incorporating the model into their practices. As necessary, the model should be reweighted according to initial results and also in light of reports from prosecutors. Indeed, testing, validating, and modifying is a central task in the field of machine learning (Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018), and it should be an essential component of the office's project. The reweighting of the model according to prosecutors' insights also is important as it would guide future data collection: the office would have a better idea as to what key points of data lead to outlier results, and they could instruct their data personnel to begin collecting these data points early in the process.

The fourth phase would involve experimentally testing the model. Over a several month trial period, a subset of prosecutors in the office would use the model, after which results between the control (no intervention) and the experimental (working with model) groups would be analyzed to determine whether the model was meaningfully improving prosecutorial fairness.

The example. To provide an example as to how AI models can be used to eliminate bias in making prosecutorial decisions, I reviewed data released by a large, urban prosecutor's office and used it to demonstrate the process and framework developed in the prior sections of this dissertation.

Collecting the data. These data included case information for felonies that were charged in the Cook County (Illinois) State's Attorney's Office from 2011-2016. Key variables included information regarding the charges (including charge severity), the date and time of the alleged

incident, the defendant's age, race, and gender, and various information about the arresting authority as well as the court authority involved (courthouse, judge, etc.).

The dataset lacked information regarding criminal history, although some information regarding criminal history could be inferred from charging inputs. This is a significant omission, given that criminal history significantly informs prosecutorial decision making. If I were to work with this district attorney's office, my first task would be to develop a plan for building out the historical record with criminal history and for storing criminal history moving forward. Neither would be overly difficult. The latter is just a matter of transcription; the former would require looking up each defendant's current criminal history and backdating it to each alleged offense appearing in the record (for example, if the record shows an arrest in May 2006, then the criminal history for that arrest would only include the items from before May 2006). This would be rather easy to represent in code. The office also failed to provide detailed descriptions of the crimes, such as those sometimes found in police reports. This would be easy to add as well.

Constructing the models. In this jurisdiction, charges have different classes. X is the most severe charge class, followed by 1 through 4, with 4 being the least severe charge class; and, also, there are misdemeanors, which are less severe than a 4. A charge reduction is when a defendant is initially charged with an offense of a certain degree but receives a plea bargain in which the most severe guilty charge is less severe than what was initially charged. For example, if a defendant is charged with a class 1 felony but receives a class 2 felony via plea bargain, then the defendant has received a charge reduction.

I chose charge reductions as the outcome variable since its usefulness to prosecutors is clear and since it is an intuitive variable. Also, charge reductions are a good choice because they represent a clear decisional point. Criminologist Besiki Kutateladze, who has worked extensively

with prosecutors' offices, has discussed the importance of prosecutorial interventions based on specific and definite decisional points (Kutateladze, 2020).

In building a model to predict charge reductions, it is tempting to assume that one can just remove the race variable and all will be well. However, this would be a significant mistake, since any included variables that are correlated with race still contain information about the protected characteristic (for example, much of the race information is likely captured by geography, which is represented by precinct, courthouse, and so on). The model then would suffer from what is called "omitted variable bias," such that race would be absent from the model but would still affect the outcomes (Clarke, 2005).

Ensemble modeling process (EMP). Next, I constructed a series of models that, taken together, represent the "Ensemble Modeling Process." The goal of EMP is to present a snapshot of possible worlds: for a new case, I show how the case would resolve if the prosecutor treated the defendant how he or she historically would have been treated. Second, I show how the case would resolve if the defendant were treated according to some definition of fairness. Third, I show how a case would resolve if the minority defendant were treated as if he or she were White. For this EMP, I decided to focus on just one crime category, one in which there was significant racial disparity: retail theft. In addition, I focused on only the two most common races listed for defendants: White and Black.

Predictive model: Historical record model. To build the purely predictive model, I prepared the dataset by identifying directly or indirectly predictive variables, removing missing values, and creating a training and test split to avoid overfitting. Next, I transformed the data using one-hot encoding (Kleppman, 2017), such that categorical variables were removed (e.g., judge name) and new binary variables reflective of those categorical variables were added. To

continue with the judge example, pursuant to one-hot encoding, instead of having a judge column with different judges' names as entries, I then had each judge's name as a separate column, with the entries being 1 (yes, judge was assigned to case) or 0 (no, judge was not assigned). In this way, the dataset was transformed into a sparse dataset, that is, one that is primarily populated by zeros.

I then chose to use the XGBoost algorithm, which has been shown to achieve state-of-the-art results on many machine learning challenges (Chen & Guestrin, 2016). It also is especially adept at handling sparse data. This technique uses a gradient boosting framework that creates models sequentially, predicting the errors of prior models until no further improvement to the prediction can be made.

As above, I trained the model using a randomly selected subset of 80% of the data, bounding the selection process so that it kept the proportion of the dependent variable (charge reduction) reflective of the overall rate. After training, the model yielded an AUC of .87, which is considered excellent (Ting Lee, et al., 2013). AUC is an abbreviation for area under the curve, where the curve is the receiver operating characteristic curve (the ROC). In short, AUC measures the two-dimensional area underneath the ROC curve. In plain English, AUC represents the probability that a random positive outcome is positioned to the right of a random negative outcome. If AUC = 1, then whenever a random positive outcome is presented, the model will position it to the right of a random negative outcomes; in other words, the model performs flawlessly. If AUC = 0, then the opposite occurs, and the model is 100% wrong. Our score of .87 represents a well-performing model.

As an intuitive indicator of model accuracy, we can compare our model performance to the baseline probabilities. For the test subset of cases, when the defendant was a Black

individual, he or she received charge reductions in 11.2% of cases. For White defendants, this was 18.3%. With our predictive model, the rates for Black individuals and White individuals, respectively, was 11.6% and 22.8% (see Figure 7.1).

Modified predictive model: "Fair" model. Second, I created a model that was bounded by a "fairness" procedure: this model was supposed to be race-neutral. In designing an algorithm, it is important to consider the context of what is being translated into code. Huq recommends using different algorithms for different crimes and depending upon the undergirding policy goal (Huq, 2019). For our model here, I used "demographic parity" (i.e., the requirement that predictions be independent of race) as the fairness definition because, when it comes predicting charge reductions, I really am predicting something that is largely independent of counterfactuals. If someone is given a class 2 felony, they will spend more time in prison than someone given a class 3 felony. There is no counterfactual. It is not fully a question of this person committing more crimes when they are released or of them not appearing for a court date.

Incapacitation and its ability to prevent future crime is one component of incarceration decisions, but it is a minor one. There also are considerations of retribution, restitution, normative validation, and much more. In essence, if the question is about anything, it is about justice: does robbery warrant 24 months in prison or does it warrant 29 months in prison? Demographic parity makes sense given this specific context. Note that the context here is very different than that of risk assessments, where discrete outcomes (e.g., recidivism, court appearance) are considered.

Even though I am using a specific type of classification system and definition of fairness, I am not recommending this definition as the one that should always be used. Rather, this is just an illustrative example, as the results make clear. There are a host of modeling techniques

available, and fairness must be determined contextually and reviewed *ex post* (Kroll et al., 2017). The particular approach a district attorney's office takes is up to the office, and I do not mean to constrain or influence such choices here.

In brief, for the race-neutral model, I used an approach that penalizes unfairness in the model training process. Specifically, I applied a method designed by Bechavod and Ligett (2018), wherein two group-dependent regularization terms are added to the loss function. These terms penalize differences in the false positive rates (FPR) and false negative rates (FNR) between groups defined by a protected variable—in this instance, race. This approach can be used with different types of models, including support vector machines (SVM) and, that which I used here, logistic regression. I used logistic regression as the base algorithm for this suggestive model because it was fully tested in Bechavod and Ligett's article. This model is the second suggestive model: it describes, in essence, how the case would resolve if the defendant were treated race-neutrally.

After training the model, I found charge reduction rates for Black and White defendants of 10.2% and 20.4% respectively (see Figure 7.1). As is evident in the figure, the ostensibly race-neutral model shifted the predictions slightly, but it did not really decrease racial disparity. Why might this be? Because the historical case record itself is biased, bounding a model so that FPR and FNR are held relatively equal across groups results in a model that is "fair," in a sense, but it is not a model that addresses the underlying racial disparities embedded in the historical record, such as those reflected in proxies for race, including a defendant's municipality or the court facility in which a case was processed. Rather, it is a model that, if used without contextual awareness and delicacy, threatens to enshrine such disparities moving forward.

Suggestive model: White transformation model. Finally, and most importantly, I created a suggestive model: for a new case, I show how that case would resolve if a minority defendant is treated as if he or she were a White defendant. To build this model, I used an approach that has been proposed in the computer science literature: alter information about the protected variables from the set of covariates to be used in the predictive models (Calders & Verwer, 2010). In other words, I did not just change the defendant's race from Black to White; I also altered covariates of Black race so that they appeared to reflect White race. Importantly, I did not alter essential and permissible features, such as information relating to the criminal offense (the crime charged, the section code of the crime, the severity of the crime, etc.). I only altered inessential features: ones that have no valid reason for impacting the outcome of a criminal case, which would include geographic features, such as court location, and incident city, and also features like the arresting officer's precinct and the judge attached to the case.

The transformation was as follows. Given a Black defendant charged with a specific criminal offense, I culled *inessential* features attached to a randomly selected White defendant charged with the same offense, and I transformed the Black defendant's features to those of the White defendant.[27] This was repeated for all of the Black defendants who had been withheld from model training (they were in the validation subset of defendants). Then I tested these transformed cases with the already trained model.

The impact of the transformation was significant: the charge reduction rate for Black defendants rose all the way to 16.7%, while for White defendants it was 15.5% (see Figure 7.1). For comparison, for this subset with outliers excluded, the actual charge reduction rate for Black

[27] There were a few criminal offenses that were rarely charged; these outliers were excluded from the suggestive model. These represent cases that, were this system in place, the machine would flag them as outliers not amenable to transformation, and prosecutors would have to analyze them without machine guidance.

defendants was 10.8%, and for White defendants it was 14.9%, and the rates for the historical model with outliers removed was 11.6% for Black defendants and 15.5% for White defendants.

Before I conclude this discussion, I must note that, as mentioned above, had the dataset contained additional case information, such as criminal history or a narrative reporting of the crime, the transformation of Black defendants to seemingly White defendants would require even more delicacy and nuance, although it does not foreclose the possibility of similar success as that achieved here.

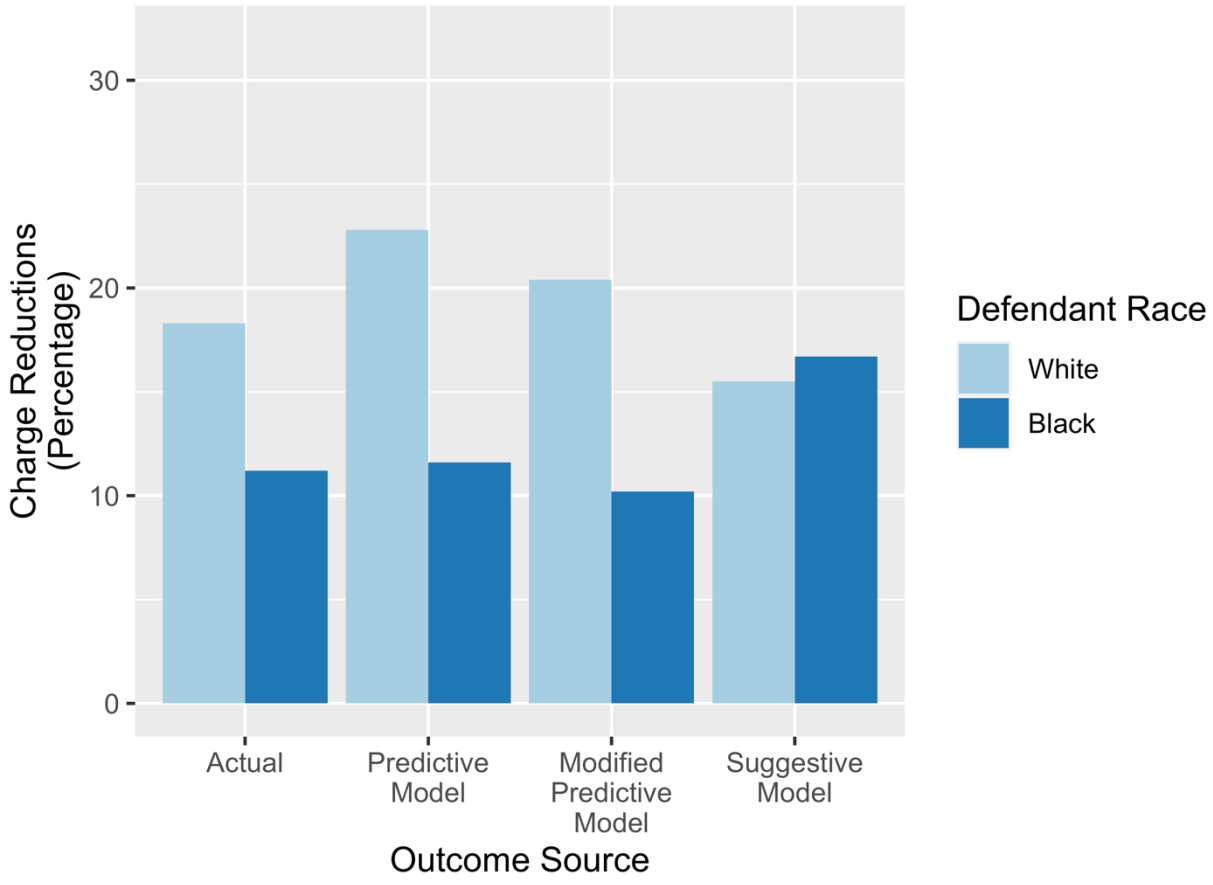


Figure 7.1 Charge reduction percentages by defendant race across the different models.

An example of the full suggestive model in place. Here is how the EMP might work for prosecutorial training and, later, actual cases. Consider the following case.

- At 3:18 pm on July 16, 2013, the defendant, a Black female who was 31 years-old at the time of the alleged incident, was arrested in Calumet City, Illinois, by the Calumet City Police Department.
- The alleged incident was a retail theft occurring that same day. The chapter, act, and section of the specific alleged crime are 720(5)§16-25(a)(1).
- The full retail value of the property stolen was less than $300. While this typically would be a misdemeanor, in this instance, given the defendant's criminal history, it was a Class 4 felony, which is punishable by one to three years in prison and fines of up to $25,000.
- The presiding court was the Markham Courthouse located in Illinois' District 6.

While the prosecutor in the district attorney's office would receive information like this to review, along with other available information (more extensive information), he or she also should receive the outputs from our models. For example, the case file would include the following visualization.

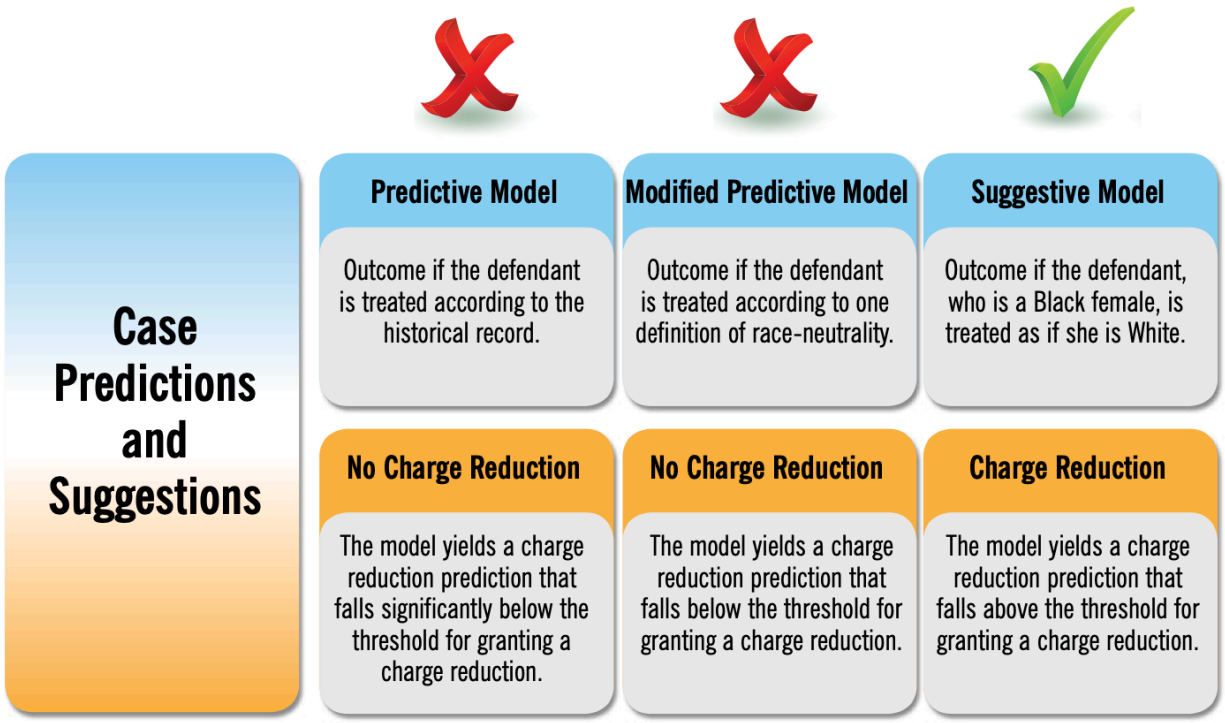


Figure 7.2 Schematic of information that, for the test case presented, could be provided to prosecutors evaluating such a case.

In actuality, and in accord with the purely predictive model, this defendant did not receive a charge reduction. The predictive model showed a very low likelihood of a charge reduction. Likewise, for the modified predictive model, the one in which defendants ostensibly are treated race-neutrally, the model showed a charge reduction likelihood of 20 percent, which fell below the threshold, and thus a charge reduction was not recommended. After I employed the procedure for changing the defendant's race to White (and simultaneously changing the covariates so that they reflected, to the machine, that the defendant was White), the new suggestive model ("Suggestive Model" in the figure above) showed a nearly certain likelihood of a charge reduction, and thus a charge reduction was recommended.

Note that neither here nor, especially, in the visualization, are the specific factors and statistics belabored. This is by design. As discussed above, prosecutors should not be burdened with interpreting statistics. What conclusions to draw from the statistics must be outlined by the leadership in a district attorney's office (e.g., leadership, working with technical experts, may decide what threshold is appropriate for recommending a charge reduction), and prosecutors should only be tasked with viewing immediately interpretable and applicable results.

Within a district attorney office's prosecutorial training sessions, this model can be used to discuss decision making. Perhaps the prosecutor believes that the defendant should not receive a charge reduction because the defendant's criminal history triggers mandatory filing behavior that does not permit reductions. (This is the type of feedback that would be used to modify the model moving forward, so that it is in accord with the office's policies.) Or, perhaps the prosecutor might believe that the defendant should not receive a charge reduction, but the prosecutor is unsure as to what led him or her to reach this conclusion. In this instance, the model might provide insight into the prosecutor's own psychology and decision making process. Regardless, once the model is sufficiently vetted in this environment, it then can be used to guide decision making on new and current cases.

A Potential Use Case of Technology to Correct for Bias in Criminal Law

Legal technology also can increase fairness in civil legal outcomes. Working on a startup that emerged from Princeton University in 2019, I have built an artificial legal intelligence for civil law firms. In brief, the AI is currently being deployed in the personal injury sector. Personal injury (PI) attorneys currently rely on a bespoke, time-intensive process of predicting case value to inform decisions on whether or not to take on new cases, as well as whether to accept or reject settlement offers. This process is far from optimal, with 50% of settlement rejections resulting in

lost revenue averaging $28,000-$43,000 per case. Access to data is a main hurdle, as data on low-level civil suits is often confidential and stored in disparate locations; these suits typically occur at the state level or are settled out-of-court and are not recorded in a centralized database, thus constraining attorneys to make decisions with few references and incomplete information available. As a result, a minority of cases contribute to a majority of law firm revenue (firms estimate that 10-20% of cases contribute to 80-90% of revenue), meaning the remaining cases occupy a disproportionate amount of time, effort, and expense. To address these issues, the startup has developed an AI system that standardizes, systematizes, and externalizes the process of civil case outcome and value prediction, providing PI attorneys with the advanced tools to optimize decision-making, allowing them to increase revenue, clients, and caseload. The solution consolidates data from across numerous law firms to train algorithms and translate learnings into computational predictions. Figure 7.3 provides an example of the output.

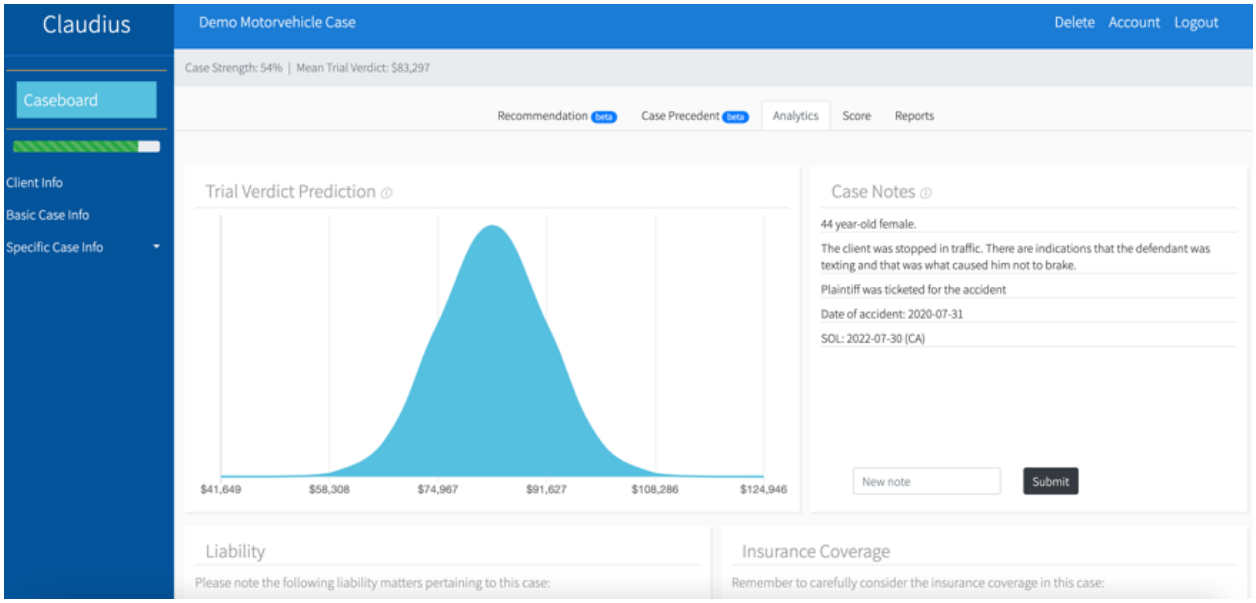


Figure 7.3 Example of output from an artificial legal intelligence developed for civil legal cases.

Even while the AI was in development, a potential case that went beyond revenue and efficiency became clear: the AI could guide decision making towards more fair and equitable outcomes. Unlike the prosecutorial example, in civil cases like this, the outcomes are largely determined not by the attorney but rather by insurance companies (as in criminal law, few civil cases make it all the way to trial, where jurors would decide). The attorneys using the AI had suspicions that the insurance companies were giving worse settlement offers to minority plaintiffs. But their suspicions were just suspicions; even the larger offices did not have the data necessary—nor the data analysts—to draw such conclusions.

Thus, it became clear that the AI could be used to inform the attorneys when their clients were receiving settlement offers that fell below what the AI predicted for a White plaintiff with a similar case. In other words, the attorney could receive a statistically-backed second opinion regarding the settlement offer, and the attorney could then use this opinion (and the underlying cases cited) to argue for more equitable settlements from the insurance companies. This is a prime example of what I have been arguing for: AI acting not as a replacement for human legal decision making but as a supplement to it, one that precipitates fair results. Moreover, it is clear that such analysis can easily be added into the output shown above; there is no requisite change in workflow. In this way, AI can be used for bias correction at the same time other goals—such as efficiency, revenue maximization—are pursued.

Conclusion

While we typically think of artificial intelligence in terms of automation (replacing human decision making), we instead should think of it as a second opinion, a guide that can raise red flags and recorrect human decisions that stray towards bias. In this chapter, I outlined a method and process for how legal technology can be used to anchor attorneys on racially fair

legal outcomes, showing what this might look like in both criminal and civil law. However, as I have pointed out, this route is not without risk. Algorithms that can help the human actors in the legal system mitigate racial bias can also exacerbate bias unless they are implemented with caution and care. Computational decision making tools need a careful framework for their development and implementation. In this chapter, I have outlined essential tasks that, if accomplished, will maximize the good and minimize the risk.

Note

The research presented in this chapter was conducted by Joel Cooper, Eric Glynn, Akbar Agha, and the dissertation author.

Concluding Thoughts

Across the three parts of this dissertation, I provided evidence of bias in key legal decision makers, I proffered a new understanding of what undergirds such bias, and I outlined a process for using artificial intelligence to help guide legal decision makers to more just and equitable outcomes. But I included a cautionary note: there are significant psychological forces at play when it comes to bias and technology, and outlining or even building a solution is not equivalent to instantiating it. With such insight, I hope this dissertation not just sets the stage for the latter but shows just how close we are to making it a reality.

References

- Ægisdóttier, S., White, M.J., Spengler, P.M., Maugherman, A.S., Anderson, L.A., Cook, R.S., Nichols, C.N., Lampropoulos, G.K., Walker, B.S., Cohen, G., & Rush, J.D. (2006). The metaanalysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34*(3), 341-382.
- Albonetti, C. A. (1990) Race and the probability of pleading guilty. *Journal of Quantitative Criminology, 6*, 315-34.
- Albonetti, C. A. (1991) An integration of theories to explain judicial discretion. *Social Problem, 38*, 247-66.
- Albonetti, C. A. (1992). Charge reduction: An analysis of prosecutorial discretion in burglary and robbery cases. *Journal of Quantitative Criminology*, *8*(3), 317-333.
- Albonetti, C. A., & Hepburn, J. R. (1996). Prosecutorial discretion to defer criminalization: The effects of defendant's ascribed and achieved status characteristics. *Journal of Quantitative Criminology*, *12*(1), 63-81.
- Alexander, M. (2010). *The new Jim Crow: Mass incarceration in the age of colorblindness*. New York: The New Press.
- Aliverti, E., Lum, K., Johndrow, J. E., Dunson, D. B. (2018). Removing the influence of a group variable in high-dimensional predictive modelling. *arXiv*. Retrieved from https://arxiv.org/abs/1810.08255
- Allport, G. W. and Postman, L. (1947). *The psychology of rumor*. New York: Henry Holt.
- Alschuler, A. W. (2003). The changing purposes of criminal punishment: A retrospective on the past century and some thoughts about the next. *The University of Chicago Law Review, 70*(1), 1-22. https://doi.org/10.2307/1600541

- Altenburger, K. M., & Ho, D. E. (2018). When algorithms import private bias into public enforcement: The promises and limitations of statistical de-biasing solutions. *Journal of Institutional and Theoretical Economics*. Retrieved from https://dho.stanford.edu/wp-content/uploads/JITE-FinalVersion.pdf
- Alter, A. L., Stern, C., Granot, Y., & Balcetis, E. (2016). The "bad is black" effect: Why people believe evildoers have darker skin than do-gooders. *Personality and Social Psychology Bulletin, 42*(12), 1653-1665. https://doi.org/10.1177/0146167216669123
- American Bar Association. (August 5, 2019). Legal profession statistics. *American Bar Association*. Retrieved from www.americanbar.org
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Amini, A., Soleimany, A., Schwarting, W., Bhatia, S., & Rus, D. (2019). Uncovering and mitigating algorithmic bias through learned latent structure. *Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence, Ethics, and Society*. Retrieved from http://www.aies-conference.com
- Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (May 23, 2016). *Machine bias. ProPublica*. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
- Arkes, H. R., & Tetlock, P. E. (2004). Attributions of implicit prejudice, or "Would Jesse Jackson 'fail' the Implicit Association Test?" *Psychological Inquiry, 15*(4), 257–278. doi: 10.1207/s15327965pli1504_01
- Arnold Foundation. (2018). *Public safety assessment: Risk factors and formula*. Retrieved from http://www.arnoldfoundation.org

- Asendorpf, J. B., Banse, R., & Mucke, D. (2002). Double dissociation between implicit and explicit personality self-concept: The case of shy behavior. *Journal of Personality & Social Psychology, 83,* 380-393. doi: 10.1037/0022-3514.83.2.380
- Association of Corporate Counsel. (2019). Global legal department benchmarking report. Retrieved from https://www.acc.com
- Avery, J. D. & Avery, J. J. (Eds.). (2019). *The stigma of addiction: An essential guide*. New York: Springer Publishing. doi: 10.1007/978-3-030-02580-9_1
- Avery, J., & Cooper, J. (Eds.). (2020a). *Bias in the law: A definitive look at racial prejudice in the U.S. criminal justice system*. Maryland: Lexington Books.
- Avery, J., & Cooper, J. (2020b). Racial bias in post-arrest and pretrial decision making: The problem and a solution. *The Cornell Journal of Law & Public Policy, 29*, 257-294.
- Avery, J., & Cooper, J. (2020c). Technology in the legal system: Uses and abuses. In J. Avery & J. Cooper (Eds.), *Bias in the law: Racial prejudice in the U.S. criminal justice system* (pp. 161-188). Maryland: Lexington Books.
- Baker, A. (2018, July 15). Confronting implicit bias in the New York Police Department. *The New York Times*. Retrieved from https://www.nytimes.com
- Baker, P. (2015, July 14). Obama calls for effort to fix a "broken system" of criminal justice. *The New York Times*. Retrieved from https://www.nytimes.com
- Balko, R. (June 10, 2020). There's overwhelming evidence that the criminal-justice system is racist. Here's the proof. *The Washington Post*. Retrieved from https://www.washingtonpost.com/graphics/2020/opinions/systemic-racism-police-evidence-criminal-justice-system/

- Banaji, M. R., Bazerman, M. H., & Chugh, D. (2003). How (un)ethical are you? *Harvard Business Review, 81*(12).
- Banse, R., Seise, J., & Zerbes, N. (2001). Implicit attitudes towards homosexuality: Reliability, validity, and controllability of the IAT. *Zeitschrift für experimentelle Psychologie*, *48*(2), 145-160.
- Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671-732.
- Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2018). Consumer-lending discrimination in the era of FinTech. Retrieved from http://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf
- Bechavod, Y., & Ligett, K. (2018). Penalizing unfairness in binary classification. Retrieved from https://arxiv.org/pdf/1707.00044.pdf
- Beck, A. J., & Harrison, P. M. (2008). Sexual victimization in local jails reported by inmates, 2007. *Bureau of Justice Statistics, U.S. Department of Justice, NCJ 221946*. Retrieved from https://www.bjs.gov/content/pub/pdf/svljri07.pdf;
- Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Medford, MA: Polity Press.
- Berdejo, C. (2018). Criminalizing race: Racial disparities in plea bargaining. *Boston College Law Review, 59*(4), 1187-1249.
- Berk, R. (2008). Forecasting methods in crime and justice. *Annual Review of Law and Social Science, 4*, 219-238.

- Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments. Fairness in Criminal Justice Risk Assessments. *Sociological Methods & Research*. doi:10.1177/0049124118782533
- Bibas, S. (2009). Prosecutorial regulation versus prosecutorial accountability. *University of Pennsylvania Law Review, 157*(4), 959-1016.
- Black, H. C., et al. (1990). *Black's Law Dictionary*. West Publishing Company.
- Blair, I. V., Judd, C. M., & Chapleau, K. M. (2004). The influence of Afrocentric facial features in criminal sentencing. *Psychological Science, 15*(10), 674-679. doi: 10.1111/j.0956-7976.2004.00739.x
- Blakemore, J. (2016). Implicit racial bias and public defenders. *Georgetown Journal of Legal Ethics, 29*, 833-848.
- Blumstein, A., & Nakamura, K. (2012, Jan. 9). Paying a price, long after the crime. The New York Times. Retrieved from https://www.nytimes.com/2012/01/10/opinion/paying-a-price-long-after-the-crime.html
- Bodenhausen, G. V. (1988). Stereotypic biases in social decision making and memory: Testing process models of stereotype use. *Journal of Personality and Social Psychology, 55*(5), 726-737. https://doi.org/10.1037/0022-3514.55.5.726
- Bodenhausen, G. V., & Wyer, R. S. (1985). Effects of stereotypes in decision making and information-processing strategies. *Journal of Personality and Social Psychology, 48*(2), 267-282. doi:10.1037/0022-3514.48.2.267
- Bonczar, T. P. (2003). Prevalence of imprisonment in the U.S. population, 1974-2001. *Bureau of Justice Statistics*. Retrieved from https://www.bjs.gov

- Booth, M. A. (2019). The road to recovery: The third circuit recognizes the importance of rehabilitative needs during sentencing in *United States v. Schonewolf. Villanova Law Review, 64*, 569-597.
- Bowers, W. J., Sandys, M., & Brewer, T. W. (2004). Crossing racial boundaries: A closer look at the roots of racial bias in capital sentencing when the defendant is Black and the victim is White. *DePaul Law Review, 53*(4), 1497-1538.
- Bowles, J. (April 14, 2019). Algorithmic accountability act targets bias in AI decision-making. *Diginomica*. Retrieved from https://diginomica.com
- Brinkman, L., Todorov, A., & Dotsch, R. (2017). Visualising mental representations: A primer on noise-based reverse correlation in social psychology. *European Review of Social Psychology, 28*(1), 333-361. https://doi.org/10.1080/10463283.2017.1381469
- Bronson, J., & Carson, E. A. (2019). *Prisoners in 2017. Bureau of Justice Statistics, 22, Table 13*. Retrieved from https://www.bjs.gov/content/pub/pdf/p17.pdf
- Bruce, P., Bruce, A., & Gedeck, P. (2017). *Practical statistics for data scientists*. CA: O'Reilly Media.
- Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1-15. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf
- Burke, K., & Leben, S. (2007). Procedural fairness: A key ingredient in public satisfaction. *Court Review, 44*, 4-25.
- Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2):277–292.

- Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI Now 2017 report. *AI Now Institute*. Retrieved from https://ainowinstitute.org
- Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology, 83*(2), 284-299.
- Carpenter, T., Pogacar, R., Pullig, C., Kouril, M., Aguilar, S. J., LaBouff, J. P., … , Chakroff, A. (2018). Survey-software Implicit Association Tests: A methodological and empirical analysis. *PsyArXiv*. doi:10.31234/osf.io/hgy3z
- Carson, E. A. (2014). Prisoners in 2013. *Bureau of Justice Statistics*. Retrieved from https://www.bjs.gov
- Carson, E. A. (2015). Prisoners in 2014. *U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics*. Retrieved from www.bjs.gov
- Carson, E. A. (2018). Prisoners in 2016. *U.S. Department of Justice, Office of Justice Programs, Bureau of Justice Statistics*. Retrieved from www.bjs.gov
- Chapman, M. V., Hall, W. J., Lee, K., Colby, R., Coyne-Beasley, T., Day, S., … Payne, K. (2018). Making a difference in medical trainees' attitudes toward Latino patients: A pilot study of an intervention to modify implicit and explicit attitudes. *Social Science & Medicine*, *199*, 202-208.
- Charlesworth, T. E. S., & Banaji, M. R. (2019). Patterns of implicit and explicit attitudes: I. Long-term change and stability from 2007 to 2016. *Psychological Science, 30*(2), 174–192. https://doi.org/10.1177/0956797618813087
- Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? Retrieved from https://arxiv.org/pdf/1805.12002.pdf

- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *arXiv*:1603.02754. doi: 10.1145/2939672.2939785
- Chi-Tung, C., Chen, C., Chih-Yuan, F., Chung-Hsien, C., Yu-Tung, W., Hsu Chih-Po, . . . Chien-Hung, L. (2020). Artificial intelligence-based education assists medical students' interpretation of hip fracture. *Insights into Imaging, 11*(1). doi: 10.1186/s13244-020-00932-0
- Chin, G. (2013). Race and the disappointing right to counsel. *The Yale Law Journal, 122*(8), 2236-2259.
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. In *Big Data, Special issue on Social and Technical Trade-Offs*. Retrieved from https://arxiv.org
- Christison, G. W., Haviland, M. G., & Riggs, M. L. (2002). The medical condition regard scale: Measuring reactions to diagnoses. *Academic Medicine, 77*, 257-262. doi: 10.1097/00001888-200203000-00017
- Citron, D. K., & Pasquale, F. (2014). The scored society: Due Process for automated predictions. *Washington Law Review, 89*, 1-33.
- Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science, 22*(4):341–352.
- Cohen, M. A. (December 20, 2018). Law is lagging digital transformation—Why it matters. *Forbes*. Retrieved from https://www.forbes.com
- Cohen, T. H., & Kyckelhahn, T. (2010). Felony defendants in large urban counties. *Bureau of Justice Statistics, US Department of Justice*.

- Cohen, T. H., & Reaves, B. A. (2006). Felony defendants in large urban counties, 2002. *U.S. Department of Justice, Bureau of Justice Statistics*. Retrieved from https://www.bjs.gov/content/pub/pdf/fdluc02.pdf.
- Colburn, A., & Melander, L. A. (2018). Beyond Black and White: An Analysis of Newspaper Representations of Alleged Criminal Offenders Based on Race and Ethnicity. *Journal of Contemporary Criminal Justice, 34*(4), 383–398. doi: 10.1177/1043986218787730
- Cook County Government. (2019). Open data. *Cook County State's Attorney*. Retrieved from https://datacatalog.cookcountyil.gov
- Copas, J. & Marshall, P. (1998) The offender group reconviction scale. *Applied Statistics, 47*, 159-71.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806).
- Correll, J., Park, B., Judd, C. M., & Wittenbrink, B. (2002). The police officer's dilemma: Using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology, 83*, 1314–1329.
- Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., & Keesee, T. (2007). Across the thin blue line: Police officers and racial bias in the decision to shoot. *Journal of Personality & Social Psychology, 92*, 1006-1009. doi: 10.1037/0022-3514.92.6.1006
- Crawford, K., & Schultz, J. (2014). Big data and due process: Toward a framework to redress predictive privacy harms. *Boston College Law Review, 55*, 93-128.

- Crootof, R. (2019). Cyborg Justice and the risk of technological-legal lock-in. *Columbia Law Review Online, 119*, 233-251.
- Cummings, M. (2004). Automation bias in intelligent time critical decision support systems. *AIAA 1st Intelligent Systems Technical Conference*, 6313.
- D'Alessio, S. J., & Stolzenberg, L. (1993). Socioeconomic status and the sentencing of the traditional offender. *Journal of Criminal Justice, 21*, 61-77. doi: 10.1016/0047-2352(93)90006-9
- DA Stat. (2019). DA Stat: Data driven decision-making. *San Francisco District Attorney*. Retrieved from https://sfdistrictattorney.org/DAStat
- Dalio, R. (2017). Principles: Life and work. New York: Simon & Schuster.
- Dasgupta, N. (2004). Implicit ingroup favoritism, outgroup favoritism, and their behavioral manifestations. *Social Justice Research, 17*(2), 143-169.
- Dastin, D. (October 9, 2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com
- Davis, A. J. (2009). *Arbitrary justice: The power of the American Prosecutor*. New York: Oxford University Press.
- Davis, A. J. (2017). The prosecution of Black men. In A. J. Davis (Ed.), *Policing the Black man: Arrest, prosecution, and imprisonment* (pp. 178-208). New York: Pantheon Books.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582. http://dx.doi.org/10.1037/0003-066X.34.7.571
- Dawes, R. M., Faust, D. & Meehl, P. E. (1989). Clinical vs. actuarial judgment. *Science, 243*, 1668-1674. doi: 10.1126/science .2648573

- DeBruine, L. (2016, October 24). Young adult composite faces. *figshare*. https://doi.org/10.6084/m9.figshare.4055130.v1
- Depew, B., Eren, O., & Mocan, N. (2017). Judges, Juveniles, and In-Group Bias. *The Journal of Law and Economics*, *60*(2), 209-239. doi:10.1086/693822
- DeSantis, A., & Kayson, W. A. (1997). Defendants' characteristics of attractiveness, race, and sex and sentencing decisions. *Psychological Reports, 81*(2), 679-683. doi: 10.2466/pr0.1997.81.2.679
- Devine, P. G. (1989). Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology, 56*(1), 5-18. https://doi.org/10.1037/0022-3514.56.1.5
- Devine, P. G., & Elliot, A. J. (1995). Are racial stereotypes really fading? The Princeton trilogy revisited. *Personality and Social Psychology Bulletin, 21*(11), 1139-1150. https://doi.org/10.1177/01461672952111002
- Devine, P. G., Forscher, P. S., Austin, A. J., & Cox, W. T. L. (2012). Long-term reduction in implicit race bias: A prejudice habit-breaking intervention. *Journal of Experimental Social Psychology*, *48*, 1267–1278.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114-26. doi: 10.1037/xge0000033
- Dipboye, R. L., Fromkin, H. L., & Wiback, K. (1975). Relative importance of applicant sex, attractiveness, and scholastic standing in evaluation of job applicant resumes. *Journal of Applied Psychology, 60*(1), 39-43. https://doi.org/10.1037/h0076352

- Donahue, L. (Jan. 3, 2018). A primer on using artificial intelligence in the legal profession. *Jolt Digest, Harvard Law School*. Retrieved from https://jolt.law.harvard.edu
- Dotsch, R., & Todorov, A. (2012). Reverse correlating social face perception. *Social Psychological and Personality Science, 3*(5), 562-571. https://doi.org/10.1177/1948550611430272
- Dotsch, R., Wigboldus, D. H., Langner, O., & van Knippenberg, A. (2008). Ethnic out-group faces are biased in the prejudiced mind. *Psychological Science, 19*(10), 978-980. https://doi.org/10.1111/j.1467-9280.2008.02186.x
- Douglass, F. (1871/1969). *Frederick Douglass, the orator: An account of his life; his eminent public services; his brilliant career as orator, selections from his speeches and writings*. J. M. Gregory, (Ed.). Chicago: Afro-Am Press.
- Dovidio, J. F., & Gaertner, S. L. (2000). Aversive racism and selection decisions: 1989 and 1999. *Psychological Science, 11*(4), 315-319. https://doi.org/10.1111/1467-9280.00262
- Dovidio, J. F., Kawakami, K., & Gaertner, S. L. (2002). Implicit and explicit prejudice and interracial interaction. *Journal of Personality and Social Psychology, 82*, 62–68. doi: 10.1037//0022-3514.82.1.62
- Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.
- Duncan, B. L. (1976). Differential social perception and attribution of intergroup violence: Testing the lower limits of stereotyping of Blacks. *Journal of Personality and Social Psychology, 34*(4), 590-598. https://doi.org/10.1037//0022-3514.34.4.590
- Durose, M. R., & Langan, P. A. (2007). Felony sentences in state courts, 2004. *Bureau of Justice Statistics*. Retrieved from https://www.bjs.gov

- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226). ACM.
- Eberhardt, J. (2019). *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. New York: Viking.
- Eberhardt, J. L., Goff, P. A., Purdie, V. J., & Davies, P. G. (2004). Seeing Black: Race, crime, and visual processing. *Journal of Personality and Social Psychology, 87*(6), 876-893. doi: 10.1037/0022-3514.87.6.876
- Eberhardt, J., & Banks, S. (2019). Implicit bias puts lives in jeopardy. Can mandatory training reduce the risk? *The Los Angeles Times*. Retrieved from https://www.latimes.com/opinion/op-ed/la-oe-eberhardt-banks-implicit-bias-training-20190712-story.html
- Edkins, V. A. (2011). Defense attorney plea recommendations and client race: Does zealous representation apply equally to all? *Law and Human Behavior, 35*, 413-425. doi: 10.1007/s10979-010-9254-0
- Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a 'right to an explanation' is probably not the remedy you are looking for. *Duke Law and Technology Review, 16*, 18-84.
- Eisenberg, T., & Johnson, S. L. (2004). Implicit racial attitudes of death penalty lawyers. *Depaul Law Review, 53*, 1539-1556.
- Erel, I., Stern, L. H., Tan, C., & Weisbach, M. S. (2018). Could machine learning help companies select better board directors. *Harvard Business Review*. Retrieved from https://hbr.org/2018/04/research-could-machine-learning-help-companies-select-better-board-directors

- Esqueda, C. W. (1997). European American students' perceptions of crimes committed by five racial groups. *Journal of Applied Social Psychology, 27*(16), 1406-1420. https://doi.org/10.1111/j.1559-1816.1997.tb01605.x
- Eubanks, V. (2017). *Automating inequality: How high-tech tools profile, police, and punish the poor*. New York: St. Martin's Press.
- European Commission (2018). Guidelines on automated individual decision-making and profiling for the purposes of Regulation 2016/679. *European Commission*. Retrieved from https://ec.europa.eu
- Everett, J. A. C., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*, 772–787. doi: 10.1037/xge0000165
- Executive Office of the President. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. *The White House*. Retrieved from https://obamawhitehouse.archives.gov
- Fazio, R. H., & Hilden, L. E. (2001). Emotional reactions to a seemingly prejudiced response: The role of automatically activated racial attitudes and motivation to control prejudiced reactions. *Personality and Social Psychology Bulletin*, *27*(5), 538-549.
- Fazio, R. H., & Olson, M. A. (2003). Implicit measures in social cognition research: Their meaning and use. *Annual review of psychology*, *54*(1), 297-327.
- Fazio, R. H., Jackson, J. R., Dunton, B. C., & Williams, C. J. (1995). Variability in automatic activation as an unobtrusive measure of racial attitudes: A bona fide pipeline? *Journal of personality and social psychology, 69*(6), 1013-1027.

- Federal Bureau of Investigations. (2019). *2018 Crime in the United States. Table 43A. Arrests by race and ethnicity.* Retrieved from https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/topic-pages/tables/table-43
- Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B. A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology, 117*(3), 522–559. doi:10.1037/pspa0000160
- Fountaine, T., McCarthy, B., & Saleh, T. (July-August, 2019). Building the AI-powered organization. *Harvard Business Review*, 62-73.
- Franklin, T. W. (2010). The intersection of defendants' race, gender, and age in prosecutorial decision making. *Journal of Criminal Justice*, *38*(2), 185-192.
- Frenzel, E. D., & Ball, J. D. (2007). Effects of individual characteristics on plea negotiations under sentencing guidelines. *Journal of Ethnicity in Criminal Justice, 5*, 59-82.
- Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2014). *An introduction to statistical learning: With applications in R*. New York: Springer.
- Garvie, C., & Frankle, J. (April 7, 2016). Facial-recognition software might have a racial bias problem. *The Atlantic*. Retrieved from https://www.theatlantic.com
- Gates, S. W., Perry, V. G., & Zorn, P. M. (2002). Automated underwriting in mortgage lending: Good news for the underserved? *Housing Policy Debate, 13*(2), 369-391.
- Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological bulletin*, *132*(5), 692.

- Gawronski, B., & De Houwer, J. (2014). Implicit measures in social and personality psychology. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 283–310). Cambridge University Press.
- Gecker, J. (June 12, 2019). San Francisco prosecutors turn to AI to reduce racial bias. *The Washington Post*. Retrieved from https://www.washingtonpost.com
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. New York: Cambridge University Press.
- Gendreau, P., Little, T. & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*, 575-607.
- Gilbert, S. A., & Johnson, M. T. (1996). The Federal Judicial Center's 1996 survey of guideline experience. *Federal Sentencing Reporter*, *9*(2), 87-93.
- Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology*, *94*(2), 292-306.
- Gordon, R. A. (1993). The effect of strong versus weak evidence on the assessment of race stereotypic and race nonstereotypic crimes. *Journal of Applied Social Psychology, 23*(9), 734-749. https://doi.org/10.1111/j.1559-1816.1993.tb01112.x
- Gordon, R. A., Bindrim, T. A., McNicholas, M. L., & Walden, T. L. (1988). Perceptions of blue-collar and white-collar crime: The effect of defendant race on simulated juror decisions. *The Journal of Social Psychology, 128*(2), 191-197. https://doi.org/10.1080/00224545.1988.9711362

- Gordon, R. A., Michels, J. L., & Nelson, C. L. (1996). Majority group perceptions of criminal behavior: The accuracy of race-related crime stereotypes. *Journal of Applied Social Psychology, 26*(2), 148-159. https://doi.org/10.1111/j.1559-1816.1996.tb01843.x
- Gottfredson, S.D., & Gottfredson, D.M. (1986). Accuracy of prediction models. In Blumstein, A. et al., (eds.), *Criminal Careers and "Career Criminals," Vol. II*. Washington, DC: National Academy of Sciences Press.
- Gramlich, J. (2019, June 11). Only 2% of federal criminal defendants go to trial, and most who do are found guilty. *Pew Research Center*. Retrieved from https://www.pewresearch.org
- Grant, P. R., & Holmes, J. G. (1981). The integration of implicit personality theory schemas and stereotype images. *Social Psychology Quarterly*, 107-115. https://doi.org/10.2307/3033706
- Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*, 619. doi: 10.1126/science.1134475
- Green, A. R., Carney, D. R., Pallin, D. J., Ngo, L. H., Raymond, K. L., Iezzoni, L. I., & Banaji, M. R. (2007). Implicit bias among physicians and its prediction of thrombolysis decisions for black and White patients. *Journal of General Internal Medicine*, *22*(9), 1231-1238. doi: 10.1007/s11606-007-0258-5
- Greenfield, A. D. (1921). Malum prohibitum. *America Bar Association Journal, 7*, 493-495.
- Greentarget. (2018). *2018 Legal industry outlook*. Retrieved from https://greentarget.com/
- Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. doi: 10.1037/0033-295X.102.1.4

- Greenwald, A. G., & Krieger, L. H. (2006). Implicit bias: Scientific foundations. *California Law Review, 94*(4), 945. doi: 10.2307/20439056
- Greenwald, A. G., Banaji, M. R., & Nosek, B. A. (2015). Statistically small effects of the Implicit Association Test can have societally large effects. *Journal of Personality and Social Psychology*, *108*(4), 553-561. doi: 10.1037/pspa0000016
- Greenwald, A. G., McGhee, D. E., & Schwartz, J. K. L. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*, 1464–1480. doi: 10.1037/0022-3514.74.6.1464
- Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*(2), 197-216. doi: 10.1037/0022-3514.85.2.197
- Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality & Social Psychology, 97*, 17-41. doi: 10.1037/a0015575
- Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323. doi: 10.1037/1076-8971.2.2.293
- Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*, 19 –30. doi: 10.1037/1040-3590.12.1.19

- Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814
- Haney, C. (2003). Mental health issues in long-term solitary and "supermax" confinement. *Crime & Delinquency*, *49*(1), 124-156.
- Hanna, M. (Feb. 6, 2018). What happened when New Jersey stopped relying on case bail? *The Philadelphia Inquirer*. Retrieved from http://www2.philly.com
- Hao, K. (Jan. 21, 2019). AI is sending people to jail—and getting it wrong. *MIT Technology Review*. Retrieved from https://www.technologyreview.com
- Harcourt, B. E. (2008). *Against prediction: Profiling, policing, and punishment in an actuarial age*. Chicago: University of Chicago Press.
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* (pp. 3315-3323).
- Hawkins, D. F. (1981). Causal attribution and punishment for crime. *Deviant Behavior, 2*, 207-230.
- Henry, P. J., & Sears, D. O. (2002). The symbolic racism 2000 scale. *Political Psychology, 23*, 253-283.
- Hetey, R. C., & Eberhardt, J. L. (2018). The numbers don't speak for themselves: Racial disparities and the persistence of inequality in the criminal justice system. *Current Directions in Psychological Science, 27*(3), 183-187. doi: 10.1177/0963721418763931
- Highhouse, S. (2008). Stubborn reliance on intuition and subjectivity in employee selection. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 1*, 333–342. doi: 10.1111/j.1754-9434.2008.00058.x

- Hill, K. (2020, Dec. 29). Another arrest, and jail time, due to a bad facial recognition match. *The New York Times*. Retrieved from https://www.nytimes.com/2020/12/29/technology/facial-recognition-misidentify-jail
- Ho, D. E. (2018). Judging statistical criticism. *Observational Studies, 4*, 42–56.
- Hofmann, W., Gawronski, B., Gschwender, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit asocial test and explicit self-report measures. *Personality and Social Psychology Bulletin, 10,* 1369-1385. doi: 10.1177/0146167205275613
- Huebner, B. M., & Bynum, T. S. (2008). The role of race and ethnicity in parole decisions. *Criminology, 46*(4), 907-938. doi: 10.1111/j.1745-9125.2008.00130.x
- Human Rights Watch (2000). Punishment and prejudice: Racial Disparities in the war on drugs. *HRW Reports, 12*(2). Retrieved from www.hrw.org
- Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal, 68*(6), 1043-1134.
- Imhoff, R., Woelki, J., Hanke, S., & Dotsch, R. (2013). Warmth and competence in your face! Visual encoding of stereotype content. *Frontiers in Psychology, 4*, 386. https://doi.org/10.3389/fpsyg.2013.00386
- Jacobs, J. B. (2015). *The eternal criminal record*. Cambridge, MA: Harvard University Press.
- Jacoby-Senghor, D. S., Sinclair, S., & Shelton, J. N. (2016). A lesson in bias: The relationship between implicit racial bias and performance in pedagogical contexts. *Journal of Experimental Social Psychology, 63*, 50–55. doi: 10.1016/j.jesp.2015.10.010

- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning: With applications in R*. New York: Springer.
- Jennings, W. G., Richards, T. N., Smith, M. D., Bjerregaard, B., & Fogel, S. J. (2014). A critical examination of the "White victim effect" and death penalty decision-making from a propensity score matching approach: The North Carolina experience. *Journal of Criminal Justice*, *42*(5), 384-398.
- Johndrow, J. E., & Lum, K. (2018). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Annals of Applied Statistics*. Retrieved from https://arxiv.org/abs/1703.04957
- Johnson, S., Blume, J. H., Eisenberg, T., Hans, V. P., & Wells, M. T. (2012). The Delaware death penalty: An empirical study. *Iowa Law Review, 97*, 1925-1964.
- Jones, C. S., & Kaplan, M. F. (2003). The effects of racially stereotypical crimes on juror decision-making and information-processing strategies. *Basic and Applied Social Psychology, 25*(1), 1-13. https://doi.org/10.1207/S15324834BASP2501_1
- Kahan, D. M. (1997). Ignorance of law in an excuse--but only for the virtuous. *Michigan Law Review, 96*(1), 127-154.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, & Giroux.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454. doi:10.1017/cbo9780511809477.004
- Kamiran, F., Žliobaitė, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems, 35*(3), 613-644.

- Kang, J., Bennett, M., Carbado, D., Casey, P., Dasgupta, N., Faigman, D., … Mnookin, J. (2012). Implicit bias in the courtroom. *UCLA Law Review, 59*, 1124-1186.
- Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: An introduction to cluster analysis*. The Wiley-Interscience Paperback Series.
- Kellough, G., & Wortley, S. (2002). Remand for plea: Bail decisions and plea bargaining as commensurate decisions. *British Journal of Criminology, 42*, 186-210.
- Kleinberg, J. Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *The American Economic Review, 105*(5), 491-495.
- Kleinberg, J. M., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. *The 8th Innovations in Theoretical Computer Science Conference*. Retrieved from https://arxiv.org/abs/1609.05807. doi:10.4230/LIPIcs.ITCS.2017.43
- Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*(1), 237-293. doi:10.1093/qje/qjx032
- Kleppmann, M. (2017). *Designing data-intensive applications: The big ideas behind reliable, scalable, and maintainable systems*. Sebastopol, CA: O'Reilly Media.
- Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review, 165*(3), 633-706.
- Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality & Social Psychology, 67*, 596-61-.

- Krueger, J., & Clement, R. W. (1994). The truly false consensus effect: An ineradicable and egocentric bias in social perception. *Journal of Personality and Social Psychology, 67*(4), 596–610.
- Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*(4), 2025-2047. https://doi.org/10.1007/s11135-011-9640-9
- Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology, 98*(6), 1060.
- Kunst, J. R., Kteily, N., & Thomsen, L. (2019). "You little creep": Evidence of blatant dehumanization of short groups. *Social Psychological and Personality Science, 10*(2), 160-171. https://doi.org/10.1177/1948550617740613
- Kutateladze, B. (2017). Tracing charge trajectories: A study of the influence of race in charge changes at case screening, arraignment, and disposition. *Criminology, 56*(1), 123-153.
- Kutateladze, B. L., Meldrum, R., Richardson, R., Stemen, D., Webster, E., Arndt, M., … Soor, S. (2018). Prosecutorial attitudes, perspectives, and priorities: Insights from the inside. *MacArthur Foundation*. Retrieved from https://caj.fiu.edu
- Kutateladze, B., Andiloro, N., & Johnson, B. (2016). Opening Pandora's Box: How does defendant race influence plea bargaining outcomes? *Justice Quarterly, 33*(3), 398-426. doi: 10.1080/07418825.2014.915340
- Kutateladze, B., Andiloro, N., Johnson, B., & Spohn, C. (2014). Cumulative disadvantage: Examining racial and ethnic disparity in prosecution and sentencing. *Criminology, 52*(3), 514-551.

- Kutateladze, B., Lawson, V., & Andiloro, N. (2015). Does evidence really matter? An Exploratory analysis of the role of evidence in plea bargaining in felony drug cases. *Law & Human Behavior, 36*(5), 431-442.
- Kutateladze, B., Lynn, V., & Liang, E. (2012). Do race and ethnicity matter in prosecution? A review of empirical studies. *Vera Institute of Justice*. Retrieved from https://www.vera.org
- Kutateldze, B. (2020). Racial and ethnic disparities in prosecution, and what can be done to change the status quo. In J. Avery & J. Cooper (Eds.), *Bias in the law: A definitive look at racial prejudice in the U.S. criminal justice system*. Maryland: Lexington Books.
- LA Times Editorial Board. (2019). The problem with LAPD's predictive policing. *The Los Angeles Times*. Retrieved from https://www.latimes.com
- Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*, 315–330.
- Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., & Nosek, B. A. (2016). Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016.
- Lake, K. (May-June 2018). Stitch Fix's CEO on selling personal style to the mass market. *Harvard Business Review*. Retrieved from https://hbr.org
- Lakkaraju, H., Kamar, E., Carauna, R., & Leskovec, J. (2019). Faithful and customizable explanations of black box models. *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. Retrieved from https://web.stanford.edu

- Lan, Z., Zhou, G., Duan, Y., & Yan, W. (2018). AI-assisted prediction on potential health risks with regular physical examination records. *2018 IEEE Third International Conference on Data Science in Cyberspace (DSC)*. doi: 10.1109/DSC.2018.00056
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., … Van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721-723. doi: 10.1126/science.1167742
- Lecher, C. (March 21, 2018). What happens when an algorithm cuts your health care? *The Verge*. Retrieved from https://www.theverge.com
- Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics, 35*, 1243–1270.
- Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies, 40*, 153–184. doi: 10.1006/ijhc.1994.1007
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 1-16. doi: 10.1177/2053951718756684
- Levinson, J. D., & Young, D. (2010). Different shades of bias: Skin tone, implicit racial bias, and judgments of ambiguous evidence. *West Virginia Law Review, 112*, 307-350.
- Levinson, J. D., Bennett, M. W., & Hioki, K. (2017). Judging implicit bias: a national empirical study of judicial stereotypes. *Florida Law Review, 69*, 63-115.
- Levinson, J. D., Cai, H., & Young, D. (2010). Guilty by implicit racial bias: The guilty/not guilty Implicit Association Test. *Ohio State Journal of Criminal Law, 8*, 187-208.

- Liang, H., Tsui, B. Y., Ni, H., Valentim, C. C. S., Baxter, S. L., Liu, G., … Xia, H. (2019). Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nature Medicine, 25*, 433-438.
- Liebman, J. S., & Clarke, P. (2011). Minority practice, majority's burden: The death penalty today. *Ohio State Journal of Criminal Law, 9*, 255-352.
- Liptak, A. (2018, Jan. 8). Death penalty case heard by racist juror is reopened by supreme court. *The New York Times*. Retrieved from https://www.nytimes.com/2018/01/08/us/politics/death-penalty-case-heard-by-racist-juror-is-reopened-by-supreme-court.html;
- Lipton, S. (2016). The mythos of model interpretability. *Proceedings of ICML 2016*. Retrieved from https://arxiv.org/pdf/1606.03490.pdf
- Lipton, Z. C., Chouldechova, A., & McAuley, J. (2017). Does mitigating ML's disparate impact require disparate treatment? arXiv preprint arXiv:1711.07076.
- Louridas, P., & Ebert, C. (2016). Machine learning. *IEEE Software, 33*(5), 110-115.
- Ludwig, A., Cohen, L., Parsons, A., & Venters, H. (2012). Injury surveillance in New York City jails. *American Journal of Public Health, 102*(6), 1108-1111.
- Lyon, A. D. (2012). Race bias and the importance of consciousness for criminal defense attorneys. *Seattle University Law Review, 35*, 755-768.
- Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The Chicago Face Database: A free stimulus set of faces and norming data. *Behavior Research Methods, 47*, 1122-1135. doi: 10.3758/s13428-014-0532-5
- Mackaay, E., & Robillard, P. (1974). Predicting judicial decisions: The nearest neighbour rule and visual representation of case patterns. *Datenverarbeitung im Recht, 3*, 302-331.

- Madhan, B., Gayathri, H., Garhnayak, L., & Naik, E.S. (2012). Dental students' regard for patients from often-stigmatized populations: Findings from an Indian dental school. *Journal of Dental Education, 76*(2), 201-217.
- Malik, A., & Tuckfield, B. (2019). *Applied unsupervised learning with R: Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA*. Packt Publishing.
- Mancuhan, K., & Clifton, C. (2014). Combating discrimination using Bayesian networks. *Artificial Intelligence and Law, 22*(2), 211-238.
- Mayson, S. G. (2019). Bias in, bias out. *The Yale Law Journal, 128*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3257004
- McConnell, A. R., & Leibold, J. M. (2001). Relations between the Implicit Association Test, explicit racial attitudes, and discriminatory behavior. *Journal of Experimental Social Psychology, 37*, 435–442. doi: 10.1006/jesp.2000.1470
- McConnell, A. R., & Leibold, J. M. (2009). Weak criticisms and selective evidence: Reply to Blanton et al. *Journal of Applied Social Psychology, 94,* 583-589. doi: 10.1037/a0014649
- McDonald, J., & Raphael, S. (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. *IssueLab*. Retrieved from https://sfdistrictattorney.org
- McKinley, J. C. (2014, July 8). Study finds racial disparity in criminal prosecutions. *The New York Times*. Retrieved from https://www.nytimes.com
- Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the literature*. Minneapolis, MN: University of Minnesota Press.

- Meier, B. P., Robinson, M. D., & Clore, G. L. (2004). Why good guys wear white: Automatic inferences about stimulus valence based on brightness. *Psychological Science, 15*(2), 82-87. https://doi.org/10.1111/j.0963-7214.2004.01502002.x
- Meltzer, E.C., Suppes, A., Burns, S., Shuman, A., Orfanos, A., Sturiano, C.V., Charney, P., & Fins, J.J. (2013). Stigmatization of substance use disorders among internal medicine residents. *Substance Abuse, 34*(4), 356-362. doi: 10.1080/08897077.2013.815143
- Mendes, W. A., Gray, H. M., Mendoza-Denton, R., Major, B., & Epel, E. S. (2007). Why egalitarianism might be good for your health. *Psychological Science*, *18*(11), 991-998. doi: 10.1111/j.1467-9280.2007.02014.x
- Merlo, J. L., Wickens, C., & Yeh, M. (2000). Effect of reliability on cue effectiveness and display signaling. In *Proceedings of the 4th Annual Army Federated Laboratory Symposium* (pp. 27–31). College Park, MD: Army Research Laboratory.
- Metz, C. (October 22, 2018). Efforts to acknowledge the risks of new A.I. Technology. *The New York Times*. Retrieved from https://www.nytimes.com
- Meyer, D. (May 8, 2018). Uber reportedly knows why its self-driving tech didn't prevent the deadly Arizona pedestrian crash. *Fortune*. Retrieved from https://fortune.com/2018/05/08/uber-autopilot-death-reason/
- Miech, R. A., Johnston, L. D., O'Malley, P. M., Bachman, J. G., & Schulenberg, J. E. (2014). *Monitoring the future: National survey results on drug use, 1975-2014*. Ann Arbor: Institute for Social Research, The University of Michigan. Retrieved from www.monitoringthefuture.org
- Miller, A. (2018). Want less-biased decisions? Use algorithms. Harvard Business Review. Retrieved from https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

- Miller, M. L. (2004). Domination and dissatisfaction: Prosecutors as sentencers. *Stanford Law Review, 56*, 1211-1276.
- Mitchell, T. L., Haw, R. M., Pfeifer, J. E., & Meissner, C. A. (2005). Racial bias in mock juror decision-making: A meta-analytic review of defendant treatment. *Law and Human Behavior, 29*(6), 621-637. https://doi.org/10.1007/s10979-005-8122-9
- Mitratech. (2018). Catching the wave: Legal technology spend at $3 billion and growing. Retrieved from http://info.mitratech.com
- Morantz, A. (2008). Mining mining data: Bringing empirical analysis to bear on the regulation of safety and health in U.S. mining. *West Virginia Law Review, 111*, 45-74.
- Moray, N. (2003). Monitoring, complacency, skepticism and eutactic behaviour. *International Journal of Industrial Ergonomics, 31*, 175–178.
- Moray, N., & Inagaki, T. (2000). Attention and complacency. *Theoretical Issues in Ergonomics Science, 1*, 354–365.
- Motivans, M. (2015). Federal justice statistics, 2012: Statistical tables. *Bureau of Justice Statistics*. Retrieved from https://www.bjs.gov/
- Munger, F. W., & Seron, C. (2017). Race, law, and inequality: Fifty years after the civil rights era. *Annual Review of Law and Social Science, 13*, 10.1-10.20. doi: 10.1146/annurev-lawsocsci-110316-113452
- Mystal, E. (June 10, 2019). Prosecutorial data collection is coming to Connecticut. *Above the Law*. Retrieved from https://abovethelaw.com
- National Highway Traffic Safety Administration (2017). *Quick facts 2016*. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812451

- Netherland, J., & Hansen, H. B. (2016). The war on drugs that wasn't: Wasted whiteness, "dirty doctors," and race in media coverage of prescription opioid misuse. *Culture, Medicine & Psychiatry, 40*, 664–686. doi: 10.1007/s11013-016-9496-5
- Nier, J. A. (2005). How dissociated are implicit and explicit racial attitudes? A bogus pipeline approach. *Group Processes & Intergroup Relations*, *8*(1), 39-52.
- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.
- Noguchi, Y. (2018, May 17). Starbucks training focuses on the evolving study of unconscious bias. *NPR*. Retrieved from https://www.npr.org/2018/05/17/611909506/starbucks-training-focuses-on-the-evolving-study-of-unconscious-bias
- Noriega-Campero, A., Bakker, M. A., Garcia-Bulle, B., & Pentland, A. (2019). Active Fairness in Algorithmic Decision Making. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*. doi:10.1145/3306618.3314277
- Nosek, B. A., Greenwald, A. G., & Banaji, M. R. (2005). Understanding and Using the Implicit Association Test: II. Method Variables and Construct Validity. *Personality and Social Psychology Bulletin*, *31*(2), 166-180. doi: 10.1177/0146167204271418
- Nosek, B. A., Smyth, F. L., Hansen, J. J., Devos, T., Lindner, N. M., Ranganath, K. A., … Banaji, M. R. (2007). Pervasiveness and correlates of implicit attitudes and stereotypes. *European Review of Social Psychology, 18*(1), 36-88. doi: 10.1080/10463280701489053
- Note. (2017). *State v. Loomis*: Wisconsin Supreme Court requires warning before use of algorithmic risk assessments in sentencing. *Harvard Law Review, 130*, 1530-1537.

- O'Dea, C. (2018, July 31). Civil rights coalition calls for end to core element of NJ bail reform. *NJ Spotlight*. Retrieved from http://www.njspotlight.com
- O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. New York: Crown.
- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447-453. doi: 10.1126/science.aax2342
- Oleson, J. C. (2011). Risk in sentencing: Constitutionally suspect variables and evidence-based sentencing. *Southern Methodist University Law Review, 64*, 1329-1402.
- Olsen, R., Courtney, L., Warnberg, C., Samuels, J. (2018). Collecting and using data for prosecutorial decisionmaking: Findings from 2018 national survey of state prosecutors' offices. *Urban Institute*. Retrieved from https://www.urban.org
- Olson, M. A., & Fazio, R. H. (2009). Implicit and explicit measures of attitudes: The perspective of the MODE model. In R. E. Petty, R. H. Fazio, & P. Briñol (Eds.), *Attitudes: Insights from the new implicit measures* (pp. 19-63). New York, NY: Psychology Press.
- Oser, C. B., Knudsen, H. K., Staton-Tindall, M., Taxman, F., & Leukefeld, C. (2009). Organizational-level correlates of the provision of detoxification services and medication-based treatments for substance abuse in correctional institutions. *Drug and alcohol dependence, 103 Suppl 1(Suppl 1)*, S73–S81.
- Ossei-Owusu, S. (2019). The sixth amendment façade: The racial evolution of the right to counsel. *University of Pennsylvania Law Review, 167*, 1161-1239.
- Oswald, F. L., Mitchell, G., Blanton, H., Jaccard, J., & Tetlock, P. E. (2015). Using the IAT to predict ethnic and racial discrimination: Small effect sizes of unknown societal

significance. *Journal of Personality and Social Psychology, 108*(4), 562-571. Doi: 10.1037/pspa0000023

- Pachankis, J. E., Hatzenbuehler, M. L., Wang, K., Burton, C. L., Crawford, F. W., Phelan, J. C., & Link, B. G. (2017). The burden of stigma on health and well-being: A taxonomy of concealment, course, disruptiveness, aesthetics, origin, and peril across 93 stigmas. *Personality and Social Psychology Bulletin*, *44*(4), 451-474. doi: 10.1177/0146167217741313
- Page, A. (2005). Batson's blind-spot: Unconscious stereotyping and the peremptory challenge. *Boston University Law Review, 85*, 155-262.
- Pantazi, A. (March 9, 2018). What makes a good prosecutor? A new study of Melissa Nelson's office hopes to find out. *The Florida Times Union*. Retrieved from https://www.jacksonville.com
- Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human factors, 52*(3), 381-410.
- Partnership on AI. (2018). *Report on algorithmic risk assessment tools in the U.S. criminal justice system*. Retrieved from https://www.partnershiponai.org/report-on-machine-learning-in-risk-assessment-tools-in-the-u-s-criminal-justice-system/
- Pasquale, F. (2016). *The black box society: The secret algorithms that control money and information*. Cambridge: Harvard University Press.
- Pastore, A. L., & Maguire, K. (Eds.). (2003). *Sourcebook of Criminal Justice Statistics*. Retrieved from http://albany.edu/sourcebook.

- Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81*, 181-192. doi: 10.1037/0022-3514.81.2.181
- Payne, B. K. (2006). Weapon bias: Split-second decisions and unintended stereotyping. *Current Directions in Psychological Science, 15*(6), 287-291. doi: 10.1111/j.1467-8721.2006.00454.x
- Payne, B. K., Lambert, A. J., & Jacoby, L. L. (2002). Best laid plans: Effects of goals on accessibility bias and cognitive control in race-based misperceptions of weapons. *Journal of Experimental Social Psychology*, *38*(4), 384-396. https://doi.org/10.1016/S0022-1031(02)00006-9
- Payne, B. K., Vuletich, H. A., & Lundberg, K. B. (2017). The bias of crowds: How implicit bias bridges personal and systemic prejudice. *Psychological Inquiry*, *28*(4), 233-248. doi: 10.1080/1047840X.2017.1335568
- Pearce, R. G. (2005). White lawyering: Rethinking race, lawyer identity, and rule of law. *Fordham Law Review, 73*, 2081-2099.
- Penner, L. A., Dovidio, J. F., Gonzalez, R., Albrecht, T. L., Chapman, R., Foster, T., … Eggly, S. (2016). The effects of oncologist implicit racial bias in racially discordant oncology interactions. *Journal of Clinical Oncology, 34*(24), 2874-2880. doi: 10.1200/JCO.2015.66.3658
- Petsko, C. D., & Bodenhausen, G. V. (2019). Race–crime congruency effects revisited: Do we take defendants' sexual orientation into account? *Social Psychological and Personality Science, 10*(1), 73-81. https://doi.org/10.1177/1948550617736111

- Picard, S., Watkins, M., Rempel, M., & Kerodal, A. (2019). Beyond the algorithm: Pretrial reform, risk assessment, and racial fairness. *Center for Court Innovation*. Retrieved from https://www.courtinnovation.org
- Pierce, G. L., & Radelet, M. L. (2011). Death sentencing in East Baton Rouge Parish, 1990-2008. *Louisiana Law Review, 71*, 647-673.
- Plackowski, E., Hull, D., & Hull, J.H. (2014). Knowledge, attitudes, and behavioral intentions toward those with Lyme Disease. *International Journal of Education and Social Science, 1*(4), 26-30.
- Plant, E. A., Butz, D. A., & Tartakovsky, M. (2008). Interethnic interactions: Expectancies, emotions, and behavioral intentions. *Group Processes & Intergroup Relations, 11*(4), 555–574. doi: 10.1177/1368430208095827
- Plenke, M. (September 9, 2015). The reason this "racist soap dispenser" doesn't work on black skin. *Mic*. Retrieved from https://mic.com
- Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy, 3*(3), 206-231.
- Quigley, W. (2012). Racism: The crime in criminal justice. *Loyola Journal of Public Interest Law, 13*, 417-426.
- Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., & Guthriet, C. (2009). Does unconscious racial bias affect trial judges? *Notre Dame Law Review, 84*, 1195-1246.
- Radelet, M. L., & Pierce, G. L. (1985). Race and prosecutorial discretion in homicide cases. *Law and Society Review, 19*, 587.

- Rasinski, K., Woll, P., & Cooke, A. (2005). Stigma and substance use disorders. In P. W. Corrigan (Ed.), *On the stigma of mental illness: Practical strategies for research and social change* (pp. 219–236). Washington, DC: American Psychological Association.
- Ready, F. (April 10, 2019). California push to automate cannabis clearances portends future of conviction relief. *LegalTechnology News*. Retrieved from https://www.law.com
- Rehavi, M. M., & Starr, S. B. (2014). Racial disparity in federal criminal sentences. *Journal of Political Economy*, *122*(6), 1320-1354.
- Reisman, D., Schultz, J., Crawford, K., & Whittaker, M. (2018). Algorithmic impact assessments: A practical framework for public agency accountability. *AI NOW*. Retrieved from https://ainowinstitute.org/aiareport2018.pdf
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Retrieved from https://arxiv.org/pdf/1602.04938v1.pdf
- Rich, J. D., Wakeman, S. E., & Dickman, S. L. (2011). Medicine and the epidemic of incarceration in the united states. *New England Journal of Medicine, 364*, 2081-2083.
- Richardson, L., & Goff, P. (2013). Implicit racial bias in public defender triage. *The Yale Law Journal, 122*, 2626-2649.
- Riddle, T., & Sinclair, S. (2019). Racial disparities in school-based disciplinary actions are associated with county-level rates of racial bias. *Proceedings of the National Academy of Sciences*, *116*(17), 8255-8260. doi: 10.1073/pnas.1808307116
- Rieland, R. (March 5, 2018). Artificial intelligence is now used to predict crime. But is it biased? *Smithsonian*. Retrieved from https://www.smithsonianmag.com
- Romm, T. (April 11, 2018). Facebook's Zuckerberg just survived 10 hours of questioning by Congress. *The Washington Post*. Retrieved from

https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/zuckerberg-facebook-hearing-congress-house-testimony/

- Rottman, J., & Kelemen, D. (2012). Aliens behaving badly: Children's acquisition of novel purity-based morals. *Cognition, 124*(3), 356-360. https://doi.org/10.1016/j.cognition.2012.06.001
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics, 20*, 53-65. https://doi.org/10.1016/0377-0427(87)90125-7
- Rush, L. (1998). Affective reactions to multiple social stigmas. *Journal of Social Psychology, 138*(4), 421–430. doi: 10.1080/00224549809600397
- Sagar, H. A., & Schofield, J. W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology, 39*(4), 590-598. https://doi.org/10.1037/0022-3514.39.4.590
- Sah, S., Robertson, C. T., & Baughman, S. B. (2015). Blinding prosecutors to defendants' race: A policy proposal to reduce unconscious bias in the criminal justice system. *Behavioral Science & Policy*, *1*(2), 69-76. doi: 10.1353/bsp.2015.0013
- Sah, S., Tannenbaum, D., Cleary, H., Feldman, Y., Glaser, J., Lerman, A., … Winship, C. (2016). Combating biased decisionmaking & promoting justice & equal treatment. *Behavioral Science & Policy*, *2*(2), 78-87. doi: 10.1353/bsp.2016.0017
- Sawyer, J., & Gampa, A. (2018). Implicit and explicit racial attitudes changed during black lives matter. *Personality and Social Psychology Bulletin, 44*(7), 1039-1059.

- Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin, 41*(8), 1147-1163. https://doi.org/10.1177/0146167215591501
- Schein, C., Ritter, R. S., & Gray, K. (2016). Harm mediates the disgust-immorality link. *Emotion, 16*(6), 862-876. https://doi.org/10.1037/emo0000167
- Schuman, H., Steeh, C., Bobo, L. D., & Krysan, M. (1997). *Racial attitudes in America: Trends and interpretations*. Cambridge, MA: Harvard University Press.
- Schütz, H., & Six, B. (1996). How strong is the relationship between prejudice and discrimination? A meta-analytic answer. *International Journal of Intercultural Relations*, *20*(3–4), 441–462. doi: 10.1016/0147-1767(96)00028-4
- Scott, M. C., & Wahl, O. F. (2011). Substance abuse stigma and discrimination among African-American male substance users. *Stigma Research and Action, 1*(1), 60–66.
- Sears, D. O., van Laar, C., Carrillo, M., & Kosterman, R. (1997). Is it really racism? The origins of white Americans' opposition to race-targeted policies. *Public Opinion Quarterly, 61*, 16–53.
- Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review, 52*, 109-195. doi:10.2139/ssrn.2819182
- Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233-242.
- Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59-68.

- Sherman, G. D., & Clore, G. L. (2009). The color of sin: White and black are perceptual symbols of moral purity and pollution. *Psychological Science, 20*(8), 1019-1025. https://doi.org/10.1111/j.1467-9280.2009.02403.x
- Shermer, L. O., & Johnson, B. D. (2010). Criminal prosecutions: Examining prosecutorial discretion and charge reductions in U.S. Federal District Courts. *Justice Quarterly*, *27*(3), 394-430.
- Silver, N. (2012). *The signal and the noise: Why so many predictions fail– but some don't*. New York, NY: Penguin Press.
- Simon, P. (2014, March). Potholes and big data: Crowdsourcing our way to better government. *Wired*. Retrieved from https://www.wired.com
- Skorinko, J. L., & Spellman, B. A. (2013). Stereotypic crimes: How group-crime associations affect memory and (sometimes) verdicts and sentencing. *Victims & Offenders, 8*(3), 278-307. https://doi.org/10.1080/15564886.2012.755140
- Smith, R. J., & Levinson, J. D. (2012). The impact of implicit racial bias on the exercise of prosecutorial discretion. *Seattle University Law Review, 35*, 795-826.
- Snook, B., Eastwood, J., Gendreau, P., Goggin, C., & Cullen, R. M. (2007). Taking stock of criminal profiling: A narrative review and meta-analysis. *Criminal Justice and Behavior, 34*(4), 437-453. https://doi.org/10.1177/0093854806296925
- Sommers, S. R., & Norton, M. I. (2007). Race-based judgments, race neutral justifications: Experimental examination of peremptory use and the Batson challenge procedure. *Law & Human Behavior, 31*, 261.
- Southall, A. (2017, Dec. 15). To curb bad verdicts, court adds lesson on racial bias for juries. *The New York Times*. Retrieved from

https://www.nytimes.com/2017/12/15/nyregion/to- curb-bad-verdicts-court-adds-lesson-on-racial-bias-for-juries.html

- Spielkamp, M. (June 12, 2017). Inspecting algorithms for bias. *MIT Technology Review*. Retrieved from https://www.technologyreview.com
- Spohn, C. A., & Homey, J. (1993). Rape law reform and the effect of victim characteristics on case processing. *Journal of Quantitative Criminology*, *9*(4), 383-409.
- Spohn, C., & Spears, J. (1996). The effect of offender and victim characteristics on sexual assault case processing decisions. *Justice Quarterly*, *13*(4), 649-679.
- Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*, 803-872.
- Steffensmeier, D., Ulmer, J., & Kramer, J. (1998). The interaction of race, gender, and age in criminal sentencing: The punishment cost of being young, Black, and male. *Criminology, 36*, 763-798.
- Stith, K. (2008). The arc of the pendulum: Judges, prosecutors, and the exercise of discretion. *The Yale Law Journal*, *117*(7), 1420-1497.
- Substance Abuse and Mental Health Services Administration (2017). *Results from the 2016 national survey on drug use and health: Detailed tables*. Retrieved from www.samhsa.gov
- Sullivan, P. (2019, Dec. 30). Bloomberg unveils plan to fight black maternal deaths. *The Hill*. Retrieved from https://thehill.com/policy/healthcare/476223-bloomberg-unveils-plan-to-fight-black-maternal-deaths
- Sunnafrank, M., & Fontes, N. E. (1983). General and crime related racial stereotypes and influence on juridic decisions. *Cornell Journal of Social Relations, 17*, 1-15.

- Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. New York: Cambridge University Press.
- Task Force on Race and the Criminal Justice System. (2012). preliminary report on race and Washington's criminal justice system. *Seattle University Law Review, 35*, 623-671.
- Tausch, N., Kenworthy, J. B., & Hewstone, M. (2007). The confirmability and disconfirmability of trait concepts revisited: Does content matter? *Journal of Personality and Social Psychology*, *92*(3), 542–556. doi: 10.1037/0022-3514.92.3.542
- The Department of the Treasury. (2009). Federal agency data mining report, 2008. *The Department of the Treasury*. Retrieved from https://www.treasury.gov
- Thiem, K. C., Neel, R., Simpson, A. J., & Todd, A. R. (2019). Are black women and girls associated with danger? Implicit racial bias at the intersection of target age and gender. *Personality and Social Psychology Bulletin, 45*(10), 1427-1439. https://doi.org/10.1177/0146167219829182
- Tiddeman, B., Burt, M., & Perrett, D. I. (2001). Prototyping and transforming facial textures for perception research. *Computer Graphics and Applications, IEEE Computer Graphics and Applications, 21*(5), 42–50. https://doi.org/10.1109/38.946630
- Ting Lee, M.L., Gail, M., Pfeiffer, R., Satten, G., Cai, T., & Gandy, A. (Eds.) (2013). *Risk assessment and evaluation of predictions*. New York: Springer Science + Business Media.
- Tollenaar, N. (2017). *Prediction modelling for population conviction data* [Dissertation]. Utrecht University Repository. Retrieved from https://www.wodc.nl/binaries/Tollenaar_thesis_tcm28-256307.PDF
- Tolnay, S. E., & Beck, E. M. (1995). *A festival of violence: An analysis of southern lynchings, 1882-1930*. Champaign, IL: University of Illinois Press.

- Towles-Schwen, T., & Fazio, R. H. (2003). Choosing social situations: The relation between automatically activated racial attitudes and anticipated comfort interacting with African Americans. *Personality and Social Psychology Bulletin, 29*(2), 170–182. doi: 10.1177/0146167202239042
- Travers, M. L. (1995). Mistake of law in *mala prohibita* crimes. *The University of Chicago Law Review, 62*(3), 1301-1332.
- Trawalter, S., Todd, A. R., Baird, A. A., & Richeson, J. A. (2008). Attending to threat: Race-based patterns of selective attention. *Journal of Experimental Social Psychology*, *44*(5), 1322-1327. doi: 10.1016/j.jesp.2008.03.006
- Turek, M. (2019). Explainable artificial intelligence. *Defense Advanced Research Projects Agency*. Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence
- United States Department of Commerce, Census Bureau. (2015). Quick facts: United States. Retrieved from www.census.gov
- United States Department of Justice. (1991). *U.S. Attorney's manual: Principles of federal prosecution*. Retrieved from https://www.justice.gov
- Varghese, R. (February 28, 2019). America's cities are running on software from the '80s. Bloomberg Businessweek. Retrieved from https://www.bloomberg.com
- Verma, S., & Rubin, J. (2018). Fairness definitions explained. *2018 ACM/IEEE International Workshop on Software Fairness*.
- Wagner, P. (2012, August 28). Incarceration is not an equal opportunity punishment. *Prison Policy Initiative*. Retrieved from www.prisonpolicy.org
- Weinhofen, H. (1928). The purpose of punishment. *Tennessee Law Review, 7*, 145-176.

- Weisberg, J. (2018, June 7). The digital poorhouse. *The New York Review of Books*, pp. 45-47.
- Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences, 114*, 11374–11379. doi: 10.1073/pnas.1704347114
- West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. *AI Now Institute*. Retrieved from https://ainowinstitute.org
- Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., … Schwartz, O. (2018). AI Now 2018 report. *AI Now Institute*. Retrieved from: https://ainowinstitute.org/AI_Now_2018_Report.html
- Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592-598.
- Wilson, H. J., & Daugherty, P. R. (July-Aug. 2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*. Retrieved from https://hbr.org
- Wilson, J. P., Hugenberg, K., & Rule, N. O. (2017). Racial bias in judgments of physical size and formidability: From size to threat. *Journal of Personality and Social Psychology*, *113*(1), 59-80.
- Wolfgang, M. E., Figlio, R. M., Tracy, P. E., & Singer, S. I. (1985). *The national survey of crime severity*. Bureau of Justice Statistics, U.S. Department of Justice.
- Word, C., Zanna, M., & Cooper, J. (1974). The nonverbal mediation of self-fulfilling prophecies in interracial interaction. *Journal of Experimental Social Psychology, 10*, 109–120. doi: 10.1016/0022-1031(74)90059-6

- World Prison Brief. (2018). Overview: United States of America. Retrieved from http://www.prisonstudies.org
- Wright, R. F. (2017). Elected prosecutors and police accountability. In A. J. Davis (Ed.), *Policing the Black man: Arrest, prosecution, and imprisonment* (pp. 234-254). New York: Pantheon Books.
- Xu, F. K., Lofaro, N., Nosek, B. A., & Greenwald, A. G. (2018). *Race IAT 2002-2017*. Retrieved from https://osf.io/52qxl/
- Yarshefsy, E. C. (2016). Duty of outrage: The defense lawyer's obligation to speak truth to power to the prosecutor and the court when the criminal justice system is unjust. *Hofstra Law Review, 44*, 1207-1226.
- Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology, 85*, 103870. doi: 10.1016/j.jesp.2019.103870
- Young, K. M., & Petersilia, J. (2016). Keep track: Surveillance, control, and the expansion of the carceral state. *Harvard Law Review, 129*, 1318-1360.
- Zatz, M. S. (2000). The convergence of race, ethnicity, gender, and class on court decisionmaking: Looking toward the 21st Century. *Criminal Justice: The National Institute of Justice Journal, 3*(1), 503-552.
- Zhang, M. (July 1, 2015). Google photos tags two African-Americans as gorillas through facial recognition software. *Forbes*. Retrieved from https://www.forbes.com

Appendix

Note. The Appendices are numbered according to the chapter to which they correspond. So we begin with Appendix II, which corresponds to Chapter 2, as there is no appendix for Chapter 1.

Appendix Number	**Page Number**
Appendix II	219
Appendix IV	225
Appendix V	226
Appendix VI	240

Appendix II

The following contains the survey measures, in the order in which they were presented. For those questions for which the results were not reported in the primary chapter text, they are reported in this Appendix.

- After accounting for the offenses of which a defendant was convicted, to what extent do you think the following, on average, play a role in judges' sentencing decisions?
    - Response scale: 1 to 5, where 1 = no role and 5 = major role
    - Options included: Defendant's criminal history; Defendant's age; Defendant's gender; Defendant's sense of contrition; Defendant's race; Whether the defendant has dependents who live with him or her; A sentencing recommendation from the prosecutor; Other. Please specify.

Table AII.1 Perceived impact of various factors on judicial sentencing decisions, where 1 is no impact and 5 is major impact.

Condition	Mean	SD	SE	95% CI
Criminal History	4.60	0.53	0.04	[4.52, 4.68]
Prosecutor's Recommendation	4.26	0.82	0.06	[4.14, 4.38]
Age	3.25	0.85	0.06	[3.13, 3.38]
Contrition	3.25	0.98	0.07	[3.11, 3.40]
Race	2.68	1.22	0.09	[2.50, 2.86]
Gender	2.68	1.02	0.08	[2.53, 2.82]
Dependents	2.45	0.86	0.06	[2.32, 2.57]

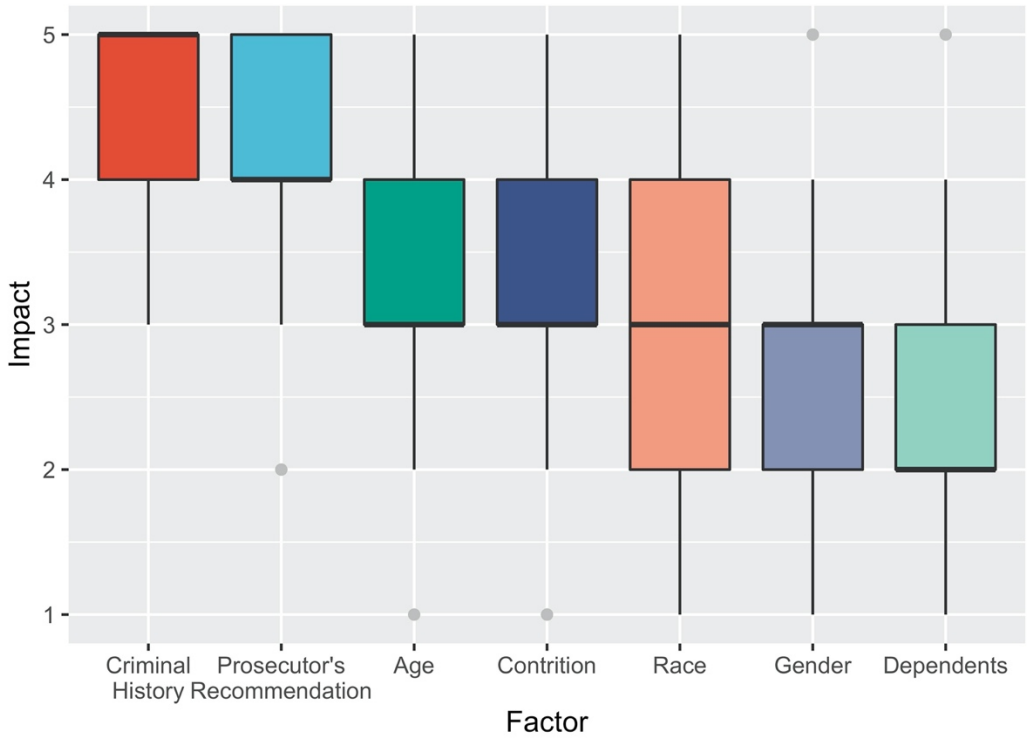


Figure AII.2 Boxplots of the perceived impact of various factors on judicial sentencing decisions. Greater scores indicate greater perceived impact.

- Implicit Association Test
- Do you think racial bias impacts your peers' legal decision making?
    - Response scale: 1 to 4, where 1 = definitely not and 4 = definitely yes
- Do you think implicit racial bias impacts your peers' legal decision making?
    - Response scale: 1 to 4, where 1 = definitely not and 4 = definitely yes
    - If they answered 3 or 4, then they were asked,
        - What's your best guess as to how often implicit racial bias impacts your peers' legal decision making? (0 = never, 100 = all the time)

- Do you think explicit racial bias impacts your peers' legal decision making?
    - Response scale: 1 to 4, where 1 = definitely not and 4 = definitely yes
    - If they answered 3 or 4, then they were asked:
        - What's your best guess as to how often explicit racial bias impacts your peers' legal decision making? (0 = never, 100 = all the time)
- To what extent do you agree with the following statement: I worry about being accused of racial bias?
    - Response scale: 1 to 5, where 1 = strongly disagree and 5 = strongly agree
- Have you ever been accused of racial bias in your professional capacity? (0 = no, 1 = yes)
    - If yes, then have the following individuals accused you of racial bias? (0 = no, 1 = yes)
        - Options included: Defendants; Other criminal defense attorneys; Prosecutors; Judges; Legislators; Researchers; Activist legal organizations; Media; Other. Please specify.
- If the goal is that every defendant will be treated fairly, how important are each of the following?
    - Response scale: 1 to 5, where 1 = not at all important and 5 = extremely important
    - Options included: Training: on implicit and explicit bias; Time: lighter dockets so each case can receive full and thoughtful treatment; Help: more support staff are needed; Safety: assurance that they can perform their work without being accused of bias; Money: better compensation to attract the best people and to help current

attorneys focus fully on the work; Information: statistical analyses that can help guide legal decision making; Other. Please specify.

- If you were tasked with eliminating racial bias in the criminal justice system, what would you do?
    - Response type: free response
- Symbolic Racism
    - It's really a matter of some people not trying hard enough; if Black people would only try harder they could be just as well off as White people.
        - Response scale: 1 to 5, where 1 = strongly disagree, 3 = neither disagree nor agree, 5 = strongly agree
    - Irish, Italian, Jewish and many other minorities overcame prejudice and worked their way up. Black people should do the same.
        - Response scale: 1 to 5, where 1 = strongly disagree, 3 = neither disagree nor agree, 5 = strongly agree
    - Some say that Black leaders have been trying to push too fast. Others feel that they haven't pushed fast enough. What do you think?
        - Response scale: 1 to 5, where 1 = going very much too slowly, 3 = moving at about the right speed, 5 = pushing very much too fast
    - How much of the racial tension that exists today do you think Black people are responsible for creating?
        - Response scale: 1 to 5, where 1 = none at all, 3 = some of it, 5 = all of it
    - How much discrimination against Black people do you feel there is today, limiting their chances to get ahead?

        - Response scale: 1 to 5, where 1 = none at all, 3 = some, 5 = a lot
    - Generations of slavery and discrimination have created conditions that make it difficult for Black people to work their way out of the lower class. [Reverse scored]
        - Response scale: 1 to 5, where 1 = strongly disagree, 3 = neither disagree nor agree, 5 = strongly agree
    - Over the past few years, Black people have gotten less than they deserve. [Reverse scored]
        - Response scale: 1 to 5, where 1 = strongly disagree, 3 = neither disagree nor agree, 5 = strongly agree
    - Over the past few years, Black people have gotten more economically than they deserve.
        - Response scale: 1 to 5, where 1 = strongly disagree, 3 = neither disagree nor agree, 5 = strongly agree
- Feeling Thermometers
    - How do you feel about White people?
        - Response scale: 0 to 100, where 0 = cold or unfavorable, 50 = no feeling at all, 100 = warm or favorable
    - How do you feel about Black people?
        - Response scale: 0 to 100, where 0 = cold or unfavorable, 50 = no feeling at all, 100 = warm or favorable

Correlation Table for Items in Study 2.2

Means, standard deviations, and correlations with confidence intervals

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. IAT	0.34	0.45											
2. Symbolic	2.10	0.85	.23** [.08, .37]										
3. Bias	2.90	0.82	-.11 [-.25, .04]	-.31** [-.44, -.17]									
4. Implicit Bias	3.15	0.79	-.17* [-.31, -.02]	-.46** [-.57, -.34]	.77** [.70, .82]								
5. Explicit Bias	2.45	0.83	-.10 [-.24, .06]	-.29** [-.42, -.14]	.60** [.50, .69]	.58** [.47, .67]							
6. Therm White	67.74	22.58	.20** [.05, .34]	.10 [-.05, .25]	-.11 [-.26, .04]	-.19* [-.33, -.04]	-.10 [-.25, .05]						
7. Therm Black	74.63	20.01	-.07 [-.22, .08]	-.29** [-.42, -.15]	.06 [-.09, .21]	.08 [-.08, .22]	.14 [-.01, .28]	.55** [.44, .65]					
8. Worry	2.59	1.23	.09 [-.06, .24]	-.15 [-.29, .00]	.22** [.07, .36]	.23** [.09, .37]	.10 [-.06, .24]	-.01 [-.16, .14]	-.06 [-.21, .09]				
9. Accused	0.31	0.46	.06 [-.09, .21]	.04 [-.12, .18]	-.03 [-.18, .12]	.04 [-.11, .19]	.00 [-.15, .15]	.13 [-.02, .27]	-.00 [-.15, .15]	.23** [.09, .37]			
10. Imp. Race	2.73	1.22	-.13 [-.28, .02]	-.39** [-.51, -.25]	.50** [.38, .61]	.42** [.28, .53]	.40** [.26, .52]	-.09 [-.24, .06]	.18* [.03, .32]	.19* [.05, .33]	-.11 [-.26, .04]		
11. Bias Imm.	2.54	1.35	.11 [-.04, .25]	.17* [.03, .32]	-.07 [-.22, .08]	-.10 [-.24, .05]	-.03 [-.18, .12]	-.04 [-.19, .11]	.11 [-.04, .26]	-.06 [-.21, .09]	-.06 [-.21, .09]	-.08 [-.23, .07]	
12. Effort	282.22	351.30	-.21** [-.35, -.06]	-.08 [-.23, .07]	.18* [.03, .32]	.12 [-.03, .26]	.08 [-.07, .23]	.02 [-.13, .17]	.09 [-.06, .24]	-.15* [-.30, -.00]	-.03 [-.18, .12]	.05 [-.10, .20]	-.13 [-.27, .02]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$. IAT = Implicit Association Test d-score. Symbolic = symbolic racism score. Bias = Do you think racial bias impacts your peers' legal decision making? Implicit Bias = same question but for implicit bias. Explicit Bias = same question but for explicit bias. Therm White = feeling thermometer towards White individuals. Therm Black = feeling thermometer towards Black individuals. Worry = To what extent do you agree with the following statement: I worry about being accused of racial bias? Accused = Have you ever been accused of racial bias in your professional capacity? Imp. Race = the impact of race: After accounting for the offenses of which a defendant was convicted, to what extent do you think defendants' race, on average, plays a role in judges' sentencing decisions? Bias Imm. = bias immunity: If the goal is that every defendant will be treated fairly, how important is the following: assurance that defense attorneys can perform their work without being accused of bias. Effort = the effort measure, that is, words written in response to the question, If you were tasked with eliminating racial bias in the criminal justice system, what would you do?

Appendix IV

Table AIV.1 Means and Standard Deviations Pertaining to Figures 4.3 and 4.4

Condition	Race Mean	Race SD	Sex Mean	Sex SD
Law Abider	6.83	2.13	8.62	1.50
Criminal	4.56	1.96	2.22	1.58
Structuring Offender	6.58	2.09	2.90	2.12
Insider Trader	8.35	1.63	1.91	1.35
Burglar	3.72	2.02	2.00	1.43
Assault Offender	3.99	2.10	2.15	1.59
Murderer	4.67	2.03	2.35	1.63
Malum In Se Offender	4.43	2.01	2.16	1.40
Malum Prohibitum Offender	6.45	2.07	2.37	1.66

Table AIV.2 Percentages Pertaining to Figure 4.5

Condition	Cluster 1 (%)	Cluster 2 (%)	Cluster 3 (%)
Law Abider	96.0	3.0	1.0
Criminal	3.9	34.9	61.2
Structuring Offender	7.8	67.8	24.4
Insider Trader	1.3	92.2	6.5
Burglar	0.7	19.9	79.5
Assault Offender	2.0	21.5	76.5
Murderer	5.9	30	64.2
Malum In Se Offender	2.6	31.6	65.8
Malum Prohibitum Offender	3.6	66.8	29.6

Appendix V

Materials for Study 5.1

Stereotypically White Crimes

- Crime Type
    - Embezzlement Condition
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. One notable group of Martians are called the Namoz: they have light green scales and purple eyes. For some reason, when Martians think of embezzlement, they think of Namoz. If one were to imagine a Namoz committing a crime, the general thinking is that the crime would be something like embezzling $500 from one's employer.
    - Tax Evasion Condition
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. One notable group of Martians are called the Namoz: they have light green scales and purple eyes. For some reason, when Martians think of tax evasion, they think of Namoz. If one were to imagine a Namoz committing a crime, the general thinking is that the crime would be something like cheating on one's Federal income tax return and avoiding paying $10,000 in taxes.
- Dependent Variables
    - Guilt (pre-conviction)

        - A Namoz was recently arrested. Allegedly, the Namoz stole a fair amount of money from a client who had entrusted the money to the Namoz's employer. There is a dispute over what precisely transpired. Regardless, the evidence presented to the grand jury apparently was sufficient, and the Namoz has been formally charged with embezzlement. The Namoz has pleaded not guilty. While the case is not a sure victory for the prosecution, it appears that the prosecution may soon be producing more and stronger evidence.
        - How do you think a Martian jury will rule on whether the defendant committed the crime? (6-point Likert scale, where 0 = definitely not guilty and 5 = definitely guilty)
    - Sentence (post-conviction)
        - A Namoz was recently arrested for stealing a fair amount of money from a client who had entrusted the money to the Namoz's employer. The evidence presented at trial apparently was convincing, and the Namoz was found guilty of embezzlement. We now are entering the sentencing phase.
        - How do you think a Martian judge will rule on how severely the defendant should be punished? (6-point Likert scale, where 0 = minimum required by law and 5 = maximum required by law)

<u>Stereotypically Black Crimes</u>

- Crime Type
    - Assault Condition

        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. One notable group of Martians are called the Namoz: they have light green scales and purple eyes. For some reason, when Martians think of assault, they think of Namoz. If one were to imagine a Namoz committing a crime, the general thinking is that the crime would be something like intentionally hitting a victim with a blunt object, such that the victim requires hospitalization.
    - Auto Theft Condition
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. One notable group of Martians are called the Namoz: they have light green scales and purple eyes. For some reason, when Martians think of auto theft, they think of Namoz. If one were to imagine a Namoz committing a crime, the general thinking is that the crime would be something like stealing a locked car and selling it.
- Dependent Variables
    - Guilt (pre-conviction)
        - A Namoz was recently arrested for assault. Allegedly, the Namoz attacked someone on the street, and the victim suffered injuries. There is a dispute over what precisely transpired. Regardless, the evidence presented to the grand jury apparently was sufficient, and the Namoz has been formally charged with assault. The Namoz has pleaded not guilty. While the case is

not a sure victory for the prosecution, it appears that the prosecution may soon be producing more and stronger evidence.

        - How do you think a Martian jury will rule on whether the defendant committed the crime? (6-point Likert scale, where 0 = definitely not guilty and 5 = definitely guilty)

- Sentence (post-conviction)
    - A Namoz was recently arrested for attacking someone on the street, and the victim suffered injuries. The evidence presented at trial apparently was convincing, and the Namoz was found guilty of assault. We now are entering the sentencing phase.
    - How do you think a Martian judge will rule on how severely the defendant should be punished? (6-point Likert scale, where 0 = minimum required by law and 5 = maximum required by law)

Additional Analyses Pertaining to Study 5.1

For the stereotypically White crimes (embezzlement vs. tax evasion), we recruited 90 participants from Prolific. After exclusions (attention check failure or non-U.S. citizen), there were 88 participants. All participants were located in the United States. Sixty percent were male, 36% were female, and the remaining selected transgender, other, or preferred not to disclose. Mean age (± SD) of the participants was 31 ± 10 years. The racial breakdown was 53% White/Caucasian, 17% Black/African American, 10% Hispanic/Latinx, 16% Asian/Pacific Islander, 1% Native American, and 2% selected "other."

Participants viewed the following two descriptions in random order:

- embezzling $500 from one's employer.
- cheating on one's Federal income tax return and avoiding paying $10,000 in taxes.

For each of the descriptions, participants indicated where the crime fell on an 11-point Likert scale in which 0 = completely moral, 5 = neither moral nor immoral, and 10 = completely immoral.

A paired two-sample t-test failed to show a difference in ratings for the two descriptions. Embezzlement (M = 7.38, SE = .18) and tax evasion (M = 7.34, SE = .21) had very similar ratings: $t(87) = .18$, $p = .86$, Cohen's $d = .02$.

For the stereotypically Black crimes (assault vs. auto theft), we recruited 90 participants from Prolific. After exclusions (attention check failure or non-U.S. citizen), there were 88 participants. All participants were located in the United States. Forty-eight percent were male, 51% were female, and the remaining selected transgender, other, or preferred not to disclose. Mean age (± SD) of the participants was 33 ± 13 years. The racial breakdown was 73%

White/Caucasian, 3% Black/African American, 8% Hispanic/Latinx, 10% Asian/Pacific Islander, 3% Middle Eastern/Arab, and 2% selected "other."

Participants viewed the following two descriptions in random order:

- intentionally hitting a victim with a blunt object, such that the victim requires hospitalization.
- stealing a locked car and selling it.

For each of the descriptions, participants indicated where the crime fell on an 11-point Likert scale in which 0 = completely moral, 5 = neither moral nor immoral, and 10 = completely immoral.

A paired two-sample t-test failed to show a difference in ratings for the two descriptions. Assault ($M = 8.31$, $SE = .20$) and auto theft ($M = 8.27$, $SE = .19$) had very similar ratings: $t(87) = .19$, $p = .85$, Cohen's $d = .02$.

Additional Figures for Study 5.1

Figure AV.1 Pre- and Post-Conviction Outcomes for the Stereotypically White Crimes

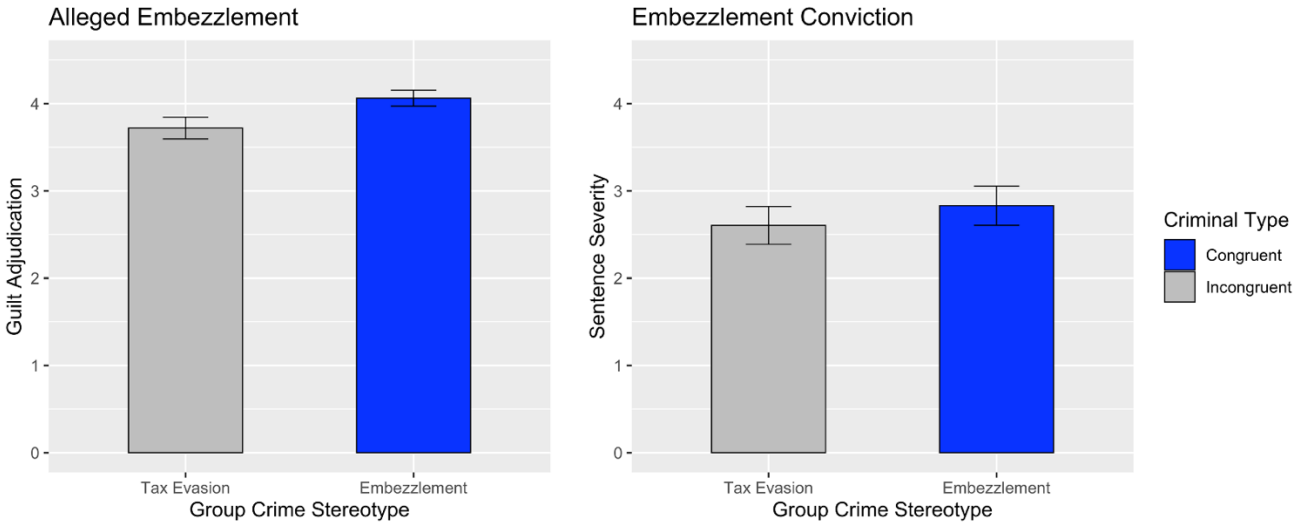
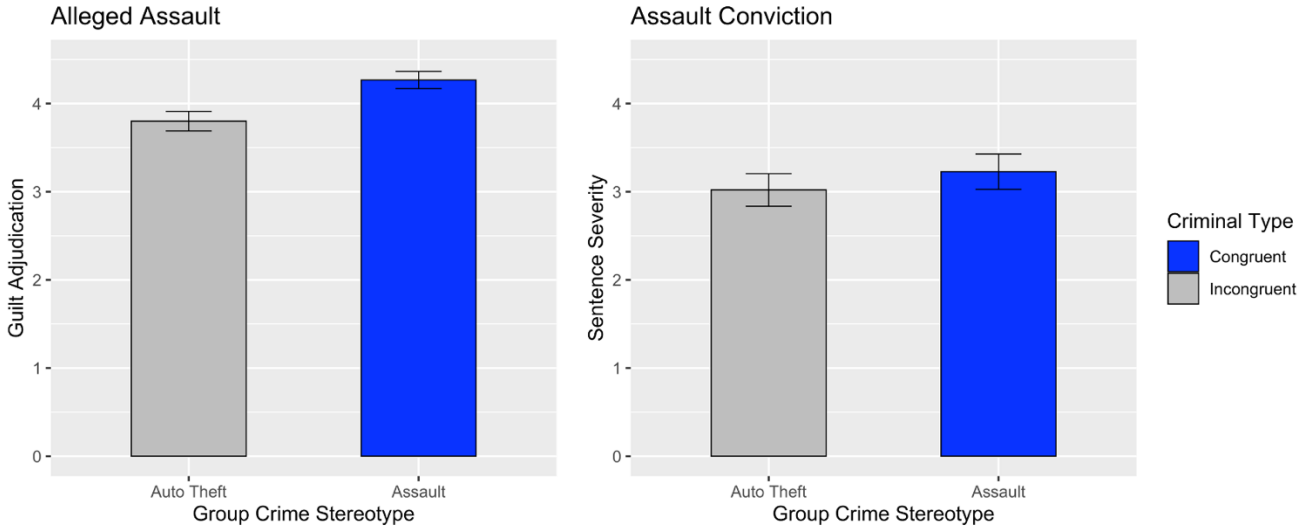


Figure AV.2 Pre- and Post-Conviction Outcomes for the Stereotypically Black Crimes



Materials for Study 5.2

- Group Stereotype
    - Moral/law-abiding
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. In other words, while there are many different groups of Martians, it doesn't take long to figure out what's up. For example, one group of Martians are called the Namoz. As soon as you see a Namoz, with their light green scales and their purple eyes, you know they belong to a group of good, moral Martians. It's not so easy to imagine a Namoz committing a crime.
    - Immoral/criminal
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. In other words, while there are many different groups of Martians, it doesn't take long to figure out what's up. For example, one group of Martians are called the Namoz. As soon as you see a Namoz, with their light green scales and their purple eyes, you know they belong to a group of bad, immoral Martians. It's easy to imagine a Namoz committing a crime.
- Dependent Variables
    - Guilt (pre-conviction)
        - A Namoz was recently arrested and charged with a serious crime. The Namoz claims innocence. While the case is not a sure victory for the

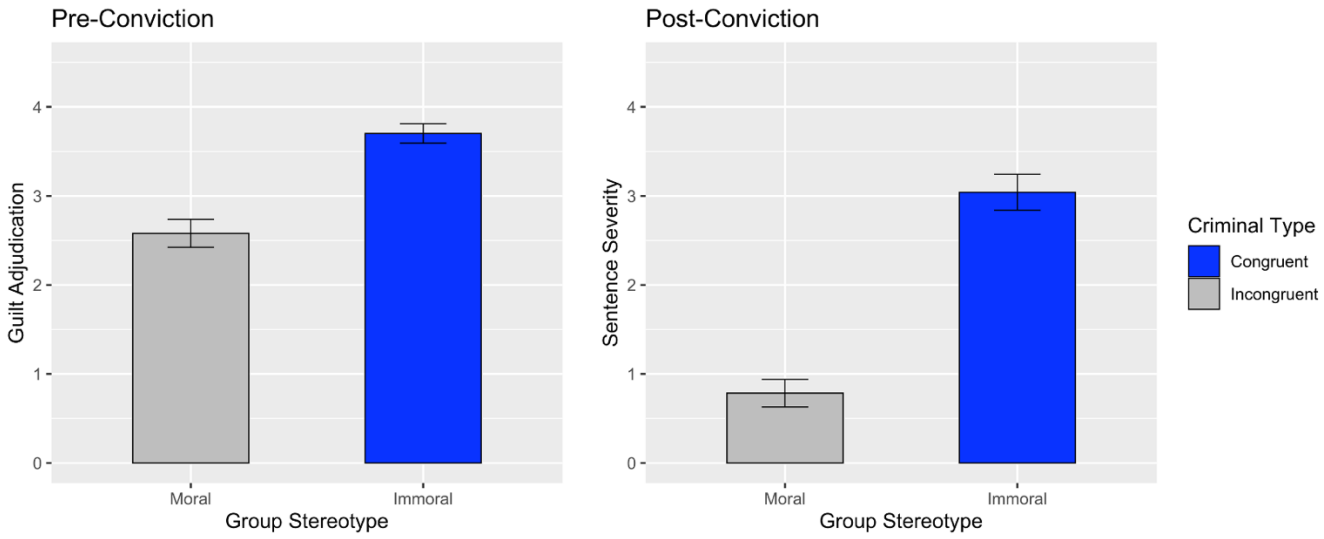
prosecution, it appears that the prosecution may soon be producing more and stronger evidence.

- How do you think a Martian jury will rule on whether the defendant committed the crime? (6-point Likert scale, where 0 = definitely not guilty and 5 = definitely guilty)

- Sentence (post-conviction)
  - A Namoz was recently arrested for a crime. The evidence presented at trial apparently was convincing, and the Namoz was found guilty. We now are entering the sentencing phase.
  - How do you think a Martian judge will rule on how severely the defendant should be punished? (6-point Likert scale, where 0 = minimum required by law and 5 = maximum required by law)

Pre- and Post-Conviction Outcomes for Stereotypically Moral/Immoral Defendants

Figure AV.3 Pre- and Post-Conviction Outcomes for Stereotypically Moral/Immoral Defendants



Materials for Study 5.3

- Group Stereotype
    - Moral/*malum prohibitum*
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. For example, one group of Martians are called the Namoz. As soon as you see a Namoz, with their light green scales and their purple eyes, you know they belong to a group of generally good Martians.
        - If one were to imagine a Namoz committing a crime, the general thinking is that the crime would surely be something like a malum prohibitum crime. Malum prohibitum means a crime that is wrong simply because it is prohibited by a statute, not because it is inherently wrong. Malum prohibitum offenses might include jaywalking, or operating a business without a license, or even some financial offenses, such as insider trading. So, for instance, if one were to imagine an insider trader, the image of a Namoz would come to mind.
    - Immoral/*malum in se*
        - Mars is a planet much like something out of Star Wars: the various groups of Martians are almost cartoonish in their traits. For example, one group of Martians are called the Namoz. As soon as you see a Namoz, with their light green scales and their purple eyes, you know they belong to a group of generally bad Martians.

        - If one were to imagine a Namoz committing a crime, the general thinking is that the crime would surely be something like a malum in se crime. Malum in se means a crime that is wrong because it is inherently wrong. Malum in se offenses might include burglary, violent crimes, and so on. So, for instance, if one were to imagine a burglar, the image of a Namoz would come to mind.

- Dependent Variables
    - Guilt (pre-conviction):
        - *Malum prohibitum* crime
            - A Namoz was just arrested for structuring (structuring is the act of breaking up financial transactions to get around federal reporting requirements). There are a few things to note about the crime. First, it isn't clear that anyone was harmed in any way; the defendant had done nothing illegal besides avoiding the federal reporting requirements. Second, there is a dispute over what precisely transpired. Regardless, the evidence presented to the grand jury apparently was sufficient, and the Namoz has been formally charged with structuring. The Namoz has pleaded not guilty. While the case is not a sure victory for the prosecution, it appears that the prosecution may soon be producing more and stronger evidence.

- How do you think a Martian jury will rule on whether the defendant committed the crime? (6-point Likert scale, where 0 = definitely not guilty and 5 = definitely guilty)

- *Malum in se* crime
  - A Namoz was recently arrested for assault and battery. There are a few things to note about the crime. First, it was a significant attack, and the victim sustained major injuries. Second, there is a dispute over what precisely transpired. Regardless, the evidence presented to the grand jury apparently was sufficient, and the Namoz has been formally charged with assault and battery. The Namoz has pleaded not guilty. While the case is not a sure victory for the prosecution, it appears that the prosecution may soon be producing more and stronger evidence.
  - How do you think a Martian jury will rule on whether the defendant committed the crime? (6-point Likert scale, where 0 = definitely not guilty and 5 = definitely guilty)

- Sentence (post-conviction)
  - *Malum prohibitum* crime
    - A Namoz was recently arrested for structuring (structuring is the act of breaking up financial transactions to get around federal reporting requirements). It was a minor offense, and it isn't clear that anyone was harmed in any way; the defendant had done nothing illegal besides avoiding the federal reporting requirements.

Regardless, the evidence presented at trial apparently was convincing, and the Namoz was found guilty. We now are entering the sentencing phase.

- How do you think a Martian judge will rule on how severely the defendant should be punished? (6-point Likert scale, where 0 = minimum required by law and 5 = maximum required by law)

- *Malum in se* crime
    - A Namoz was recently arrested for assault and battery. It was a significant attack, and the victim sustained major injuries. The evidence presented at trial apparently was convincing, and the Namoz was found guilty. We now are entering the sentencing phase.
    - How do you think a Martian judge will rule on how severely the defendant should be punished? (6-point Likert scale, where 0 = minimum required by law and 5 = maximum required by law)

Appendix VI

Materials for Study 6.1

- Prompt for Human Adjudicator Condition
    - **Overview of the Plea Bargaining Process**
    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works: When a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached.
    - But here's a key point: prosecutors generally have an idea as to what the case outcome should be. And the prosecutor negotiates with this outcome in mind. If the defendant ultimately rejects this outcome, then the case will go to trial (this does not happen often).
    - **Training**
    - Prosecutors in Cooke County go through an extensive training program designed to ensure fair and just outcomes.
- Prompt for Machine Adjudicator Condition
    - **Overview of the Plea Bargaining Process**
    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works: When a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached.

- **Machine Decision Making**
- But here's a key point: in Cooke County, there is an algorithm that was built to determine case outcomes. This machine (the algorithm) tells the prosecutor what the outcome should be. And the prosecutor negotiates with this machine-generated outcome in mind. If the defendant ultimately rejects this outcome, then the case will go to trial (this does not happen often).
- **Training**
- The machine (the algorithm) underwent an extensive creation, training, and validation program designed to ensure fair and just outcomes.

- Dependent variable
  - To what extent do you agree or disagree with the following statement? I trust the performance of the prosecutors [machine (the algorithm)] in Cooke County.
    - Graded on a 6-point Likert scale as follows: 0 = No, definitely disagree; 1= Disagree; 2 = Probably disagree; 3 = Probably agree; 4 = Agree; 5 = Yes, definitely agree

Materials for Study 6.2

- Prompt for Human Adjudicator Condition
  - **Overview of the Plea Bargaining Process**
  - In Cooke County, nearly every criminal case resolves by plea bargaining. The Cooke County prosecutors handle the negotiations. They make offers, formal requests for additional information, and so on. The prosecutors determine the final plea bargain the defendant receives.

    - Prosecutors in Cooke County go through an extensive training program designed to ensure fair and just outcomes.
    - **Evaluating Performance**
    - The office may have a problem. The evidence is preliminary, but it appears that the prosecutors are giving White defendants plea bargains that are approximately 7% better than those given to similarly-situated Black defendants.
- Prompt for Machine Adjudicator Condition
    - **Overview of the Plea Bargaining Process**
    - In Cooke County, nearly every criminal case resolves by plea bargaining. The Cooke County Prosecutor's Office has developed an algorithm (and an AI machine interface) for managing plea bargaining. The machine makes offers, formal requests for additional information, and so on. The defendant and the defense attorney respond directly to the machine. This machine actually determines the final plea bargain the defendant receives.
    - The machine (the algorithm) underwent an extensive creation, training, and validation program designed to ensure fair and just outcomes.
    - **Evaluating Performance**
    - The office may have a problem. The evidence is preliminary, but it appears that the machine is giving White defendants plea bargains that are approximately 7% better than those given to similarly-situated Black defendants.
- Dependent Variable
    - Machine version: There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop

processing cases (the office can't process cases without the machine, so they would have to postpone everything) until these disparities are eliminated?

- Human Version: There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (they would have to postpone everything) until these disparities are eliminated?
- Graded on a 6-point Likert scale as follows: 0 = Definitely no; 1= No; 2 = Probably no; 3 = Probably yes; 4 = Yes; 5 = Definitely yes

Materials for Study 6.3

- Prompt for Human Adjudicator Condition
  - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works: when a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached.
  - Prosecutors determine what the case outcome should be. They run the negotiation with this outcome as the baseline. If the defendant ultimately rejects this outcome, then the case will go to trial (this does not happen often).
  - While prosecutors in Cooke County go through an extensive training program designed to ensure fair and just outcomes, they are not perfect. They make errors. And they show some bias in favor of White defendants.

    - Specifically, a recent review of the office found that the prosecutors were giving White defendants plea bargains that were approximately 7% better than those given to similarly-situated Black defendants.
- Prompt for Machine Adjudicator Condition
    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works: when a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached.
    - In Cooke County, however, there is an algorithm that was built to determine what the case outcome should be. The machine (the algorithm) runs the negotiation with this outcome as the baseline. If the defendant ultimately rejects this outcome, then the case will go to trial (this does not happen often).
    - While the machine (the algorithm) underwent an extensive training program designed to ensure fair and just outcomes, it is not perfect. It makes errors. And it shows some bias in favor of White defendants.
    - Specifically, a recent review of the office found that the machine was giving White defendants plea bargains that were approximately 7% better than those given to similarly-situated Black defendants.
- Dependent Variables
    - Note. DV1 was presented first, then DVs 2 and 3 were presented in random order, counterbalanced across the total number of participants.
    - DV1

        - Indicate the extent to which you agree or disagree with the following statement.
        - Machine Version: I trust the performance of the machine (the algorithm) in Cooke County.
        - Human Version: I trust the performance of the prosecutors in Cooke County.
        - Graded on a 6-point Likert scale as follows: 0 = No, definitely disagree; 1 = Disagree; 2 = Probably disagree; 3 = Probably agree; 4 = Agree; 5 = Yes, definitely agree
    - DV2
        - The office does not have any funds budgeted to investigating this issue. However, it could shift funding away from other worthy causes in order to pursue this. How much funding should be redeployed so that this issue can be investigated further?
        - Graded on a 6-point Likert scale as follows: 0 = None; 1 = A very small amount; 2 = Some but not much; 3 = A moderate amount; 4 = A significant amount; 5 = As much as possible
    - DV3
        - Machine Version: There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (the office can't process cases without the machine, so they would have to postpone everything) until these disparities are eliminated?

        - Human Version: There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (they would have to postpone everything) until these disparities are eliminated?
        - Graded on a 6-point Likert scale as follows: 0 = Definitely No; 1 = No; 2 = Probably No; 3 = Probably Yes; 4 = Yes; 5 = Definitely Yes

Materials for Study 6.4a

- Prompt for Human Adjudicator Condition
    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works: when a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached. In short, the prosecutor controls this process and determines the final plea bargain the defendant receives.
    - New prosecutors in Cooke County go through an extensive training program designed to ensure fair and just outcomes. The training program, which is run in collaboration with a local university, is considered state-of-the-art.
- Prompt for Machine Adjudicator Condition
    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works in most counties: when a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others,

negotiating with the defendant until a plea deal is reached. In short, the prosecutor controls this process and determines the final plea bargain the defendant receives.

    - However, in Cooke County, there is an algorithm that was built to determine case outcomes. This machine (the algorithm) says what the outcome should be. Instead of the prosecutor, it is the machine that controls the plea bargaining process.
    - The machine (the algorithm) was created through an extensive training program designed to ensure fair and just outcomes. The training program, which was run in collaboration with a local university, is considered state-of-the-art.
    - After training, the machine was tested against the performance of actual prosecutors in the office. Its performance was equal to the average performance of the prosecutors. In other words, this is a machine that is as good at being a prosecutor as an average prosecutor is.

- Dependent Variable
    - Indicate the extent to which you agree or disagree with the following statement.
    - Machine Version: I trust the performance of the machine (the algorithm) in Cooke County.
    - Human Version: I trust the performance of the prosecutors in Cooke County.
    - Graded on a 6-point Likert scale as follows: 0 = No, definitely disagree; 1 = Disagree; 2 = Probably disagree; 3 = Probably agree; 4 = Agree; 5 = Yes, definitely agree

Materials for Study 6.4b

- Prompt for Human Adjudicator Condition

    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works: when a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached. In short, the prosecutor controls this process and determines the final plea bargain the defendant receives.
    - New prosecutors in Cooke County go through an extensive training program designed to ensure fair and just outcomes. The training program, which is run in collaboration with a local university, is considered state-of-the-art.
    - In spite of their training, Cooke County prosecutors are not perfect. They make errors. And they show some bias in favor of White defendants. Specifically, a recent review of the office found that the prosecutors were giving White defendants plea bargains that were approximately 6% better than those given to similarly-situated Black defendants.
- Prompt for Machine Adjudicator Condition
    - In Cooke County, nearly every criminal case resolves by plea bargaining. Here's how plea bargaining works in most counties: when a defendant is arrested, the prosecutor decides which offenses to charge the defendant with. Over the course of the case, the prosecutor will decide to drop some charges and keep others, negotiating with the defendant until a plea deal is reached. In short, the prosecutor controls this process and determines the final plea bargain the defendant receives.

- However, in Cooke County, there is an algorithm that was built to determine case outcomes. This machine (the algorithm) says what the outcome should be. Instead of the prosecutor, it is the machine that controls the plea bargaining process.
- The machine (the algorithm) was created through an extensive training program designed to ensure fair and just outcomes. The training program, which was run in collaboration with a local university, is considered state-of-the-art.
- After training, the machine was tested against the performance of actual prosecutors in the office. Its performance was equal to the average performance of the prosecutors. In other words, this is a machine that is as good at being a prosecutor as an average prosecutor is.
- In spite of its training, the machine is not perfect. It makes errors. And it shows some bias in favor of White defendants. Specifically, a recent review of the office found that the machine was giving White defendants plea bargains that were approximately 6% better than those given to similarly-situated Black defendants.

- Dependent Variables
  - DV1
    - Indicate the extent to which you agree or disagree with the following statement.
    - Machine Version: I trust the performance of the machine (the algorithm) in Cooke County.
    - Human Version: I trust the performance of the prosecutors in Cooke County.

        - Graded on a 6-point Likert scale as follows: 0 = No, definitely disagree; 1 = Disagree; 2 = Probably disagree; 3 = Probably agree; 4 = Agree; 5 = Yes, definitely agree
    - DV2
        - Machine Version: There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (the office can't process cases without the machine, so they would have to postpone everything) until these disparities are eliminated?
        - Human Version: There has been some discussion over whether the office should take an extremely drastic measure. What do you think: should the office stop processing cases (they would have to postpone everything) until these disparities are eliminated?
        - Graded on a 6-point Likert scale as follows: 0 = Definitely No; 1 = No; 2 = Probably No; 3 = Probably Yes; 4 = Yes; 5 = Definitely Yes

Materials for Study 6.4c

The design of Study 6.4c is identical to Study 6.4b, except that the percentages for the rates of bias are different. In Study 6.4c, the human prosecutor bias was specified as follows: "Specifically, a recent review of the office found that the prosecutors were giving White defendants plea bargains that were approximately 8% better than those given to similarly-situated Black defendants." The machine prosecutor bias was specified as follows: "Specifically,

a recent review of the office found that the machine was giving White defendants plea bargains that were approximately 4% better than those given to similarly-situated Black defendants."

Materials for Study 6.5

Study 6.5 made use of the same design as Study 6.4a. The first question asked of participants was the trust item, which was analyzed and discussed in Study 6.4a. After the trust item, participants answered 9 items. These items were presented in random order, except for the last item, which addressed a hypothetical and needed to be situated last. The items were graded on a 6-point Likert scale as follows: 0 = No, definitely disagree; 1 = Disagree; 2 = Probably disagree; 3 = Probably agree; 4 = Agree; 5 = Yes, definitely agree. The items were as follows:

- On the following pages, you will see a series of statements. Indicate the extent to which you agree or disagree with them.
    - DV One
        - Human
            - I worry about data privacy and security given that defendants' information is processed by the Cooke County Prosecutor's Office.
        - Machine
            - I worry about data privacy and security given that defendants' information is processed by the machine (the algorithm) in Cooke County.
    - DV Two
        - Human

        - I think it's just weird and uncomfortable to have to deal with a prosecutor.
    - Machine
        - I think it's just weird and uncomfortable to have to deal with the machine rather than a real human prosecutor.
- DV Three
    - Human
        - I think the process is flawed: every situation is unique, and prosecutors can't possibly take into account all possible scenarios, such as intent, coercion, and so on.
    - Machine
        - I think the process is flawed: every situation is unique, and the machine can't possibly take into account all possible scenarios, such as intent, coercion, and so on.
- DV Four
    - Human
        - I think the process is flawed: every person is unique, and prosecutors can't possibly take into account individual differences, such as whether the defendant is a good person who just made a mistake.
    - Machine
        - I think the process is flawed: every person is unique, and the machine can't possibly take into account individual differences,

such as whether the defendant is a good person who just made a mistake.

- DV Five
  - Human
    - I think the prosecutors will discriminate, and they will not be fair.
  - Machine
    - I think the machine will discriminate, and it will not be fair.
- DV Six
  - Human
    - I think this sounds like a system designed by privileged White people in order to perpetuate, rather than diminish, racial bias.
  - Machine
    - I think this sounds like an algorithm designed by privileged White people in order to perpetuate, rather than diminish, racial bias.
- DV Seven
  - Human
    - Prosecutors have the moral capacity necessary for the job they are performing.
  - Machine
    - The machine has the moral capacity necessary for the job it is performing.
- DV Eight
  - Human

- Thinking about the plea bargaining process in Cooke County, I am worried that the prosecutors are too inflexible and that this will create a barrier to change and future improvement in criminal justice.

- Machine
  - Thinking about the plea bargaining process in Cooke County, I am worried that the machine is too inflexible and that this will create a barrier to change and future improvement in criminal justice.

- DV Nine
  - Human
    - Now imagine that you have been arrested in Cooke County. To what extent do you agree with the following statement? I think I will get from the prosecutor a result that's better than what other defendants in my position would receive.
  - Machine
    - Now imagine that you have been arrested in Cooke County. To what extent do you agree with the following statement? I think I will get from the machine a result that's better than what other defendants in my position would receive.

Supplemental Table for Study 6.5

Main Effects for the Independent Variable ("Trust") and the Dependent Variables by Participant Race

		Black		White		
		Mean	SD	Mean	SD	p_{adj}
Trust	Total	2.65	1.27	2.56	1.16	1.00
	Human	2.85	1.03	2.54	1.09	1.00
	Machine	2.42	1.47	2.57	1.24	1.00
Privacy	Total	2.89	1.35	2.71	1.39	1.00
	Human	2.81	1.20	2.44	1.34	1.00
	Machine	2.98	1.50	2.96	1.40	1.00
Unease	Total	2.91	1.37	3.16	1.35	1.00
	Human	2.35	1.23	2.65	1.36	1.00
	Machine	3.51	1.27	3.65	1.16	1.00
Situational nuance	Total	3.62	1.20	3.47	1.19	1.00
	Human	3.56	1.18	3.17	1.19	1.00
	Machine	3.69	1.22	3.76	1.12	1.00
Defendants' uniqueness	Total	3.38	1.23	3.29	1.26	1.00
	Human	3.25	1.21	3.08	1.37	1.00
	Machine	3.51	1.25	3.49	1.12	1.00
Systemic racism	Total	2.83	1.22	2.67	1.36	1.00
	Human	2.96	1.03	3.04	1.27	1.00
	Machine	2.69	1.40	2.31	1.36	1.00
Expected discrimination	Total	2.92	1.43	2.47	1.64	.45
	Human	2.94	1.31	2.71	1.79	1.00
	Machine	2.91	1.56	2.25	1.47	.37
Moral capacity	Total	2.04	1.29	1.98	1.45	1.00
	Human	2.65	1.02	2.77	0.99	1.00
	Machine	1.40	1.25	1.24	1.42	1.00
Fear of lock-in	Total	3.26	1.14	3.39	1.24	1.00
	Human	3.04	1.13	3.15	1.32	1.00
	Machine	3.49	1.12	3.63	1.13	1.00
Self-advantage	Total	2.14	1.18	2.42	1.21	1.00
	Human	2.17	1.23	2.92	1.05	.02
	Machine	2.11	1.13	1.96	1.18	1.00

Note. Black and White refer to participants' race/ethnicity (Black/African American and White/Caucasian, respectively). Human = human adjudicator, that is, prosecutors. Machine = machine adjudicator. Total = results collapsed across human and machine. p_{adj} = two-tailed t-test with Bonferroni corrections.