DECEMBER 2020

A Research Working Paper

Using Artificial Intelligence to Improve the Fairness and Equity of Government Decision Making

Joseph J. Avery

Akbar A. Agha

Eric J. Glynn

Joel Cooper

PPCL

The Princeton Project in Computational Law

Using Artificial Intelligence to Improve the Fairness and Equity of Government Decision Making

Joseph J. Avery

Princeton University
The Princeton Project in Computational Law

Akbar A. Agha

The Princeton Project in Computational Law

Eric J. Glynn

Princeton University
The Princeton Project in Computational Law

Joel Cooper

Princeton University

TABLE OF CONTENTS

EXECUTIVE SUMMARY

Can artificial intelligence (AI) technology be used to augment high-stakes decisions in the public sector, such as in making prosecutorial charging or sentence decisions.

Numerous academic and civil rights advocates have expressed wariness about unintentional bias when relying on AI to support decisions affecting civil liberties.

Evidence shows that these kinds of decisions, when based solely on human judgement, have often resulted in biased outcomes—either intentional or unintentional. This has resulted in historically biased outcomes data, and critics are concerned that using these data to develop AI-based computational models to augment human decisions will reinforce past biases.

This report focuses specifically on such cases. We posit that rather than a liability, potentially biased historical data might be used by government agencies to improve human decisions and result in greater fairness in outcomes and organizational efficiency.

Specifically, agencies should establish a framework whereby government executives can anchor their decision making on race-neutral, machine-generated outcomes. As a proof of concept, we use prosecutorial decision making by local district attorneys to demonstrate a framework that can equip government executives with AI-augmented advanced analytics to guide their decision making in ways that reduce bias and improve outcomes.

We believe this framework can be adapted to the implementation of AI-augmented decision making in other policy domains as well.

Designing a Framework for Using AI to Augment Human Decisions.

A three-step framework for using artificial intelligence to augment human decisions based on potentially biased historical data is informed by two guiding principles—transparency and accountability. Decisions made using advanced analytics must be transparent and explainable, and advanced analytics must include strong accountability for their use, such as a right of appeal.

The first step in adopting the use of advanced analytics supported by AI is to collect relevant data. The immediate goal is to collect the data that are essential to the purpose of the organization's AI initiative.

The second step is to build computational models that reflect key values, such as fair outcomes. For example, when correcting for racial bias, the overarching objective should be to build a model that can anchor decision making on fair outcomes.

The third step is to manage the human-computer interaction. An agency's leadership will need to work with staff to change their mindset about bias from a focus on finding or not finding disparities in outcomes, to a focus on finding ways to correct such disparities by augmenting their decisions with AI-supported analytics.

Implementation Recommendations.

Agency leaders should use the following guiding principles and steps to implement a framework that relies on AI-based technologies to improve decision-making in their programs, including reducing bias.

Recommendations to Implement the Guiding Principles

Transparency: Ensure decisions made using advanced analytics are transparent and explainable. Specifically, agency leaders should inform the public that algorithmic decision making is being used to augment human decisions. Ideally, both the data and algorithmic code should be made publicly available. The data may need to be partially scrubbed and redacted for privacy reasons, but as much as possible should be disclosed, and the same is true for code.

Accountability: Ensure advanced analytics include strong accountability for their use. Specifically, agency leaders should set clear metrics for success and assess progress on those metrics in regular reports, similar to those used in financial accounting. In addition, agency leaders should establish a right of appeal that tracks the right that attaches to human decisions of a similar type. At a minimum, complainants should be allowed to contest inaccuracies in machine inputs and to present mitigating information.

Recommendations to Implement the Framework

1 Step One: Collect relevant data.

The immediate goal should be for agency leaders to collect the data that are essential to the purpose of their organization's AI initiative. Initial data collection does not need to be exhaustive. The ultimate goal is to contribute to centralized and standardized data collection that can be made available for public consumption.

2 Step Two: Build computational models that reflect key values, such as fair outcomes.

Agency leaders need to articulate, up front, the values that their decision process should reflect. Specifically, when correcting for racial bias, the overarching objective should be to build a model that can anchor decision making on fair outcomes. This can be facilitated by setting clear objectives and, when predictive analytics are employed, clearly defining what is being predicted. Ideally, this would be done via two teams of experts. One team would be comprised of technical experts to handle model construction. The other team would be comprised of subject matter experts that ensure that the models are built in accordance with their intended design.

3 Step Three: Manage the human-computer interaction.

Once the computational models are built, agency leaders need to work with staff to ensure the models are used, and used appropriately. For example, they should draft a "reasoned rule," a written account of an agency's agreed upon decision making protocol, which can be compared against both human and machine decisions. They should also introduce computational models during training sessions with hypothetical cases, so that both human and machine decisions can be vetted, critiqued, and modified. And finally, agency leaders will need to work with staff to change their mindset from a focus on experience-driven decision making to a focus on data-driven decision making.

Applying the Framework in a District Attorney's Office.

To demonstrate how AI models could be used to augment prosecutorial decisions, we reviewed data released by a large, urban prosecutor's office and used it to provide a proof of concept of the process and framework developed in this report. Working with felony theft charges, we built models that predict three possible outcomes: how a case would resolve if the prosecutor treated the defendant (1) how he or she historically would have been treated, (2) according to one definition of fairness, and (3) as if he or she were a member of the majority racial group.

When reviewing individual cases, these models can be used to provide different outcomes by which, during training sessions, prosecutors can explore their case intuitions. In addition, this example shows how the framework might be used to build a system that can preemptively anchor prosecutors on more fair and just outcomes.

Conclusion

More broadly, we conclude that if government agencies apply this framework and approach to augmenting decisions with AI and analytics—especially if they do so in concert with academics, researchers, community members, key stakeholders, and other knowledgeable parties in government, such as agencies' Chief Data Officers—they will increase that chance of success in decreasing disparities in their agency's decisions and outcomes.

INTRODUCTION

Business analysts at Gartner predict that by 2021, organizations worldwide will create $2.9 billion in business value by harnessing the use of artificial intelligence (AI) to augment human decision making.[1]

Government agencies are increasingly using AI, as well. For example, the American AI Initiative, launched in February 2019 pursuant to a presidential executive order,[2] directs federal agencies to prioritize artificial intelligence (AI) research and development and create guidelines to implement AI technology, including its use in government decision making.

At the same, there has been general concern about the use of AI in government decision making because of the fear of bias.[3] For example, the AI Now Institute at New York University focuses much of its resources on identifying unintentional bias in artificial intelligence as well as when and how such bias potentially affects human rights and liberties.[4] The Institute has partnered with the American Civil Liberties Union and has published a number of papers expressing wariness of the increasing reliance on this technology in the United States.[5]

Indeed, this wariness seems well-founded. Consider the experience of T. J. Fitzpatrick, a guest at a science fiction convention in Atlanta, Georgia in the summer of 2017.[6] Mr. Fitzpatrick, who is Black, used the sink in a bathroom at a Marriott Hotel. He placed his hands under the soap dispenser and received no soap. He assumed it was not working—until a White friend of his 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 hands but would not dispense soap for Black hands.

This story perfectly captures the fear about AI. The dispenser was using near-infrared technology, which requires that hands presented to receive soap reflect light back to it. Hands with darker skin pigments reflect less light, and so the dispenser failed to work for most Black individuals.[7] While this is not bias in the senses of racial animus or intentional prejudice, it does reflect how machines are bound by rigid rules, and these rules may unintentionally lead to unfair outcomes. This is true for automatic soap dispensers, and it is true for algorithms that predict legal case outcomes.

However, racial bias and racially disparate outcomes are not endemic to the computer era. In the criminal justice realm, racial bias goes back centuries. In 1871, Frederick Douglass, speaking to an audience that included 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.”[8] This is not surprising since psychology has shown us that bias is a human problem, one that is not necessarily diminished by experience, age, or introspection. Indeed, given that racial disparities in incarceration were extreme 150 years ago and remain extreme today, it seems likely that, unless there is a significant intervention, nothing will change.

One potentially significant way to remove racial bias is through the use of AI. Instead of shying away from data and artificial intelligence, government decision makers should embrace them and marshal them for good. We propose that government leaders use AI or advanced computational and statistical techniques to train models that—drawing on their agencies’ own historical data—enable decision makers to anchor themselves on racially fair outcomes. Just

as hospitals and other organizations are using AI to decrease variance, thus regularizing and optimizing behavior in order to meet performance and profit goals, the same can be done when the goal is racial fairness.

We acknowledge that this approach is somewhat counterintuitive. While bias is a human problem, in some instances the use of AI may exacerbate this problem. This is something we discuss later in this report. Thus, in order to design a system that uses AI to correct for bias, one must proceed with rigor and nuance. Indeed, in this report, we unpack one way in which to make this counterintuitive stance a reality.

Technological intervention is especially appealing because human biases are relatively difficult to identify and even harder to address. In contrast, machine bias is potentially knowable: computational purposes and processes can be declared prior to use and also verified once in use.[9] Even so, it appears inevitable that technology will increasingly influence government decision making. However, both public servants and lay persons are uncertain of how best to move forward. In its 2019 report, New York City's Automated Decision Systems Task Force recommended focusing on this very conundrum: while the city has machine tools that show promise for improving New Yorkers' lives, it needs to focus on how to identify and eliminate bias in those tools.[10]

This report, which was supported by a stipend from the IBM Center for The Business of Government, addresses this very point. Unless AI systems are designed and implemented correctly, they may perpetuate or even exacerbate the problems they were designed to solve. Thus, we offer two guiding principles for the design of an AI-assisted decision making framework and three steps for building such a framework. To demonstrate this framework in action, we show how this approach could be used by prosecutors in district attorney offices to reduce racial bias in their decisions. However, the general framework we developed can be applied in other government functions, as well.

The two guiding principles—transparency and accountability—must inform all aspects of artificial intelligence use. The three steps for building a framework outline the specifics that should be considered in (1) collecting the data, (2) constructing the computational models, and (3) guiding the human-computer interface. In sum, our framework shows how linking human and artificial intelligence can help remove biases in decision making in order to improve government effectiveness and fairness. While we focus on a particular use case—prosecutorial decisions by district attorneys—the lessons we draw and the principles we outline are applicable to the wider expanse of government decision making.

KEY TERMS USED IN THIS REPORT

In this report, we are focused on techniques for dealing with data. These techniques are manifold, and they arise from different fields, such as statistics, applied mathematics, computer science, and the social sciences. For reference, we clarify a few of the key terms here:

Advanced Analytics is the broadest term we use, and we use it to signify processes for discovering and interpreting meaningful patterns in data. This includes **Modeling**, a term we use to describe the use of mathematics and computer science to simulate and study the behavior of complex systems, processes, and decisions. It also includes the related but more specific concept of **Cognitive Computing**, which is the use of models to simulate human thought processes in complex situations. Advanced analytics includes predictive analytics as well, which is something on which we focus in the latter part of this report. **Predictive Analytics** is the use of various computational techniques to analyze current and historical facts and make predictions about future or counterfactual outcomes.

Artificial Intelligence is somewhat difficult to define, given the philosophical struggle to define intelligence, but we can think of it as computer systems capable of performing tasks that normally require human intelligence. Among a host of other things, it includes **Natural Language Processing**, which is AI applied to the task of processing, deciphering, and making use of human languages. It also includes **Machine Learning**, which is the use of algorithms and statistical models to detect patterns and make inferences in order to complete tasks that seemingly would require human intelligence.

1

An Overview of Removing Bias in Decision Making

"Bias" refers to a statistic that is systematically different from the population parameter being estimated. For example, *selection bias* is when certain individuals are more likely to be selected for a study than others, resulting in a biased sample. In illustrating what bias is, it is common to show a target at which multiple arrows have been shot, and all of the arrows are lodged a foot or so to the right of the bullseye. They are systematically off-target.

"*Cognitive bias*" takes the same idea and applies it to human thought. In addition to it being a statistical term, it is also a term used by behavioral scientists.[11] It refers to a systematic error in thinking that affects decision making. Countless cognitive biases have been identified:

- Some relate to memory: the serial-position effect refers to our bias of having best recall of the first and last items in a series.
- Some cognitive biases relate to attention: people are drawn to details that support things they already believe (confirmation bias; congruence bias; experimenter's bias).
- A number of cognitive biases have to do with groups. People tend to favor people or objects with which they are familiar (halo effect; in-group bias).
- Further, when people lack information, they tend to fill in characteristics based on stereotypes and prior experiences with members of different groups. These types of biases can result in significant differences in outcomes, such as disparities in earnings between men and women.

Cognitive biases can exist either implicitly or explicitly. Broadly, *explicit bias* refers to consciously held attitudes, while *implicit bias* refers to unconscious attitudes.

In this report, we focus on disparate outcomes that result from *racial bias*. We address those situations in which similarly situated individuals are treated differently based on race. Even though our specific focus is on racial bias, this framework may apply to other forms of bias, as the framework is transferable. That said, an important aspect of bias is the context in which it arises, and thus any application of such a framework requires careful understanding of the specific context that pertains to the type of bias.

Bias in the Criminal Justice System

For decades, racial bias in government decision making has been alleged—and largely proven in statistical analysis, such as the well-known report documenting bias in disability benefits determinations.[12] It is unsurprising that, to the present day, many individuals from racial minority groups distrust government decision making.[13] But in no sector of government are racial disparities as pronounced and as well-documented as in the criminal justice system.[14] African-Americans are dramatically overrepresented in U.S. prisons and jails.[15] In 2015, the U.S. population was 13 percent Black,[16] while the U.S. state prison population was 38 percent Black.[17] Wagner found that Black individuals are more than four times as likely to be incarcerated as White individuals.[18]

This report focuses on one particular governmental actor: district attorneys. District attorneys arguably have the most control over racial disparities in the criminal justice system. According to the Bureau of Justice Statistics, for every 100 felony defendants processed in urban courts, only three are convicted at trial.[19] In the federal system, only 2 percent 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 percent.[20] Thus, most cases are resolved through plea bargaining, a process

over which district attorneys have significant discretion.[21] Research has found that district attorneys are less likely to offer Black defendants 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.[22] Defendants who are Black, young, and male fare especially poorly.[23]

One reason for these racial disparities in prosecutorial decision making appears to be a lack of clear baselines. In estimating the final disposition of a case, prosecutors have little on which to base their estimations. New cases constantly arrive, and prosecutors must process these cases quickly and efficiently, all while receiving information secondhand: determining what happened and when is a matter of cobbling together reports from victims, witnesses, police officers, and investigators. In addition, prosecutors tend to rely on their own past experiences, a reliance that hazards a number of deficiencies, including the risk of small sample size bias. Given these constraints, prosecutors are especially liable to overreliance on stereotypes, such as those that attach to Black individuals.[24]

The Limited Effectiveness of Bias Interventions

A range of interventions for reducing bias has been developed and tried, but they have been largely unsuccessful. The most widely used interventions are informed by intergroup contact theory.[25] These interventions typically involve participants interacting with "out-group" members under optimal conditions, such as when individuals from different racial groups have common goals, are focused on cooperation, and so on.[26] Some of the different intergroup approaches have achieved short-term success, at least in terms of lower stated preference for one's in-group.[27] When it comes to changing something beyond stated preferences, such as actual feelings and cognitions, the results are even worse,[28] even though many types of interventions have been tried.[29]

A recent meta-analysis found that, while some changes in implicit bias can be achieved through interventions (at least in the short term), nearly all interventions have limited success.[30] Of those that show promise, the change in implicit bias is generally small, and any subsequent change in behavior is negligible.[31] Long-term success is most important, and here there is scant evidence that interventions work at all.[32] A 2016 study[33] reviewed nine interventions that had showed short-term efficacy, finding that for none of the nine interventions did the positive effects persist for more than a few days.

There is little reason to think that criminal justice interventions will be any more successful than interventions in other domains. A 2017 law review article [34] recommended hiring a more diverse pool of assistant district attorneys, but studies show that negative in-group bias might be just as great, if not greater, than negative out-group bias.[35] In 2017, three academic researchers[36] provided 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 given prison sentences rather than probation.

The New York Times and Seymour W. James Jr., the Attorney-in-Chief at the Legal Aid Society in New York City, among others, have recommended implicit bias trainings for prosecutors.[37] A 2010 *Harvard Law Review* article[38] recommended adding a discussion of implicit bias to jury instructions. But bias trainings and education alone tend not to override stereotypes. For example:

- A 2014 study by a pair of psychologists[39] showed White voters in California photographs of incarcerated people in which the racial makeup of the inmates was either 45 percent or 25 percent Black. When a greater percentage of the inmates were Black, the participants were significantly less likely to sign a real petition aimed at lessening racial inequality in prison sentencing.
- Similarly, when White New York residents read that New York's prison population was 60 percent Black (vs. 40 percent Black), they were less likely to support a petition to end a racially discriminatory NYC policing policy.[40]
- In addition, two political scientists in 2014[41] found that, when White participants were informed about racial disparities in executions, 52 percent favored the death penalty compared with 36 percent when the participants were not informed of the racial disparities.

In short, exposure to disparities may cause individuals to become more supportive of the policies that created those disparities.

Another commonly recommended solution is to blind criminal justice decision makers to defendants' race.[42] This practice is currently being implemented by the San Francisco District Attorney's Office.[43] 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.[44] In addition, prosecutors must see defendants at some point, so blinding can only impact initial case decision making.

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,[45] and they provide a basis for explaining behavior.[46] In short, individuals use non-racial stereotypes and associations to help navigate the world, and it makes sense that they also would use racial stereotypes and associations for the same.

The Promise of Big Data and AI to Reduce Bias

Research has shown that some individuals are increasingly hiding their explicit racial biases while neglecting to address more subtle forms.[47] This is especially true in the public servant context, where explicit racial bias by state employees triggers Equal Protection concerns. Thus, individuals working in public settings are relatively unlikely to reveal explicit racial bias, and interventions are forced to focus on what is less conscious and less visible: implicit bias. But implicit bias interventions, as discussed in the preceding section, have shown extremely limited efficacy. In light of this,

- We recommend a change of emphasis in government organizations, such as district attorney offices: rather than focus on implicit bias trainings, racial bias education, and similar interventions, emphasis should be placed on using technology to correct for bias.
- In addition, we recommend using technology—such as big data, artificial intelligence and predictive analytics—to form clear, unbiased baselines on which human decision makers can anchor their decisions.

In this way, government leaders can start with the problem—disparate outcomes—and once the problem is solved, then work backwards to address the deep and entrenched precipitating factors, such as in-group and out-group attitudes.

Big data, predictive analytics, and artificial intelligence have already shown tremendous promise,[48] both at the local government level,[49] and at the federal level, where there has been a long and concerted effort to determine how to build and use AI tools.[50] Such tools have been deployed in taxation,[51] health and safety,[52] and even benefits determinations. For example, the Social Security Administration's (SSA) disability system handles a heavy volume of applications per year, and it relies on more than 1,000 human adjudicators. Research has documented racial disparities in benefits awards, with Black individuals receiving less favorable outcomes.[53] To increase efficiency and decrease bias, the SSA has, in recent years, begun to use artificial intelligence in the form of machine-guided adjudication.[54] Similarly, the Securities and Exchange Commission is also using AI, but it is using it to increase detection of suspicious activity and automate the initiation of enforcement investigations.[55]

AI has also shown promise for improving the administration of criminal justice.[56] Analytics-based risk assessment has been used in pretrial detention,[57] sentencing,[58] parole,[59] and probation.[60] There are a number of examples of existing technological applications that are having positive impacts on legal processes and outcomes, including decreasing bias. For example:

- Two California counties automated the clearance of more than 50,000 marijuana convictions that had been made eligible to be cleared pursuant to a recently enacted proposition.[61] An algorithm was used to comb through government records to identify eligible individuals and file motions to clear. The automation made it so that the person with the criminal record did not have to file a petition, hire an attorney, or attend a hearing—all of those tasks were automated.[62] Because Black individuals were overrepresented in this population and stood to benefit the most from clearance of marijuana convictions, the technology worked to correct for historical bias.[63]
- The Data-Driven Justice Initiative is another good example. Johnson County, Kansas officials partnered with academics from the University of Chicago to use machine learning to divert low-level offenders from jail to mental-health services.[64] Using data from three distinct sources—emergency medical services, the mental health center, and the county's court and corrections database—the team was able to identify low-level offenders who were likely to be re-arrested and who would benefit from mental health treatment.

There are a number of other examples of government organizations using AI to correct for bias in legal proceedings:

- In San Francisco, AI is being used to blind police reports so that prosecutors do not know defendants' race.[65]
- Jurisdictions across the country are using a risk assessment tool that was designed to reduce bias in judicial decision making. In assessing dangerousness, the Public Safety Assessment Tool (PSA), which was created from an analysis of 1.5 million criminal cases from 300 U.S. jurisdictions, takes into account age and history of criminal convictions, but it omits data on race, gender, where a defendant lives, and history of criminal arrests, and other potentially discriminatory factors.[66]
- A team of economists in 2018[67] conducted groundbreaking theoretical work showing that machine predictions can lead to drastically improved legal outcomes. Specifically, these economists focused on bail decisions: those instances when judges decide to jail or release defendants based on their predictions concerning what the defendants will do if released. The algorithm built by the economists resulted in jailing defendants at existing rates while 25 percent fewer crimes were committed by individuals released on bail. Alternatively, the algorithm could be used to keep crime rates the same (that is, there would be the same number of offenses committed by released individuals) while reducing jail populations by 42 percent (significantly more defendants were released on bail). In short, compared with the human judges, the algorithm was able to reduce crime, reduce incarceration rates, or even do both.

Worries about Big Data and AI Exacerbating Bias

In spite of the promise, there are significant worries about big data, predictive analytics, and AI.[68] During President Obama's second term, the White House[69] and the U.S. Federal Trade Commission[70] issued reports on the possibility of big data analysis leading to racial discrimination. Before discussing specific instances of bias when using technology, it is worth describing how such bias emerges.

Machine learning, a technology that makes use of big data, falls within the scope of AI and powers a significant portion of predictive analytics. Generally defined, machine learning is 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 before.[71] Machine learning can be affected by a number of different biases, such as:

- **Sampling bias.** First, there is sampling bias: the population used to train the model may not accurately represent the population with which the model will interact. In fact, this is a major reason why facial-recognition systems have struggled: they are trained primarily on White faces, but they must perform on faces of every race/ethnicity.[72]
- **Base rate bias.** There also is bias that emerges from base rates. For example, construction workers may statistically be more likely to be male, but a machine that concludes that all construction workers are male would be making a significant error. And this is the type of error that machines often make: base rates are enshrined as if they are reflective of causal relationships, even when this may not be the case at all. For example, consider when Amazon attempted to automate its resume-screening procedures. Given that many of the company's early, successful employees had been male, the resume-screening machine wound up screening out nearly all female candidates.[73] The base rate (long-term, successful Amazon employees are overwhelmingly male) resulted in the machine inferring that maleness was necessary for long-term success at the company.

Different base rates across different legally protected groups are one of the thornier problems that beset fairness in predictive analytics.[74] Different groups nearly always have different base rates, and different base rates, absent perfect prediction, lead to different error rates (one can think of false positives and false negatives). In short, groups wind up being treated differently. This is a primary element of what has been called the "group fairness impossibility theorem."[75] Research in algorithmic fairness is progressing well, with significant recent breakthroughs being made,[76] but that does not diminish the importance of understanding that there are fairness constraints[77] that may need to be resolved at the policy level, as discussed later in this report.

- **Historical data bias**. Lastly, and perhaps most importantly, the primary driver of bias in technology, especially in predictive analytics, may be biases that are enshrined in the historical record. For example, if prosecutors consistently offered Black individuals worse plea bargains than White defendants, then a predictive model trained on such data would encode the bias as weights in the model. In other words, anything that indicated that a defendant was Black would be used by the machine as an indicator that the plea bargain offered should be more severe.

Applied technology has shown how these biases may play out in practice. In recent years, there have been instances of technology leading to racially biased outcomes.[78] In the introduction to this report, we mentioned how soap dispensers were dispensing soap to White individuals only.[79] Technology has often struggled with recognition problems: a number of facial-recognition systems have been shown to misidentify non-White faces at elevated rates.[80] Amazon's facial recognition technology has repeatedly been accused of racial bias.[81] These missteps have special significance in the context of this report, given that many municipal governments are considering or have implemented facial recognition technology—and with predictably mixed results.[82] This underscores the importance of continued research in this area, especially before widespread implementation.

A notable example of the impact of bias when relying on historical data comes from the healthcare system. Researchers evaluated a widely used algorithm that was tasked with predicting which patients would need additional treatment, and they found significant racial disparities.[83] These disparities were driven by the fact that the algorithm had been programed to use healthcare costs as a proxy for the health of the patient; in other words, the engineers had assumed that, in the past, those patients whose healthcare had resulted in higher costs were likely sicker and in greater need of additional care. While this was a convenient and seemingly reasonable proxy, it was flawed for one reason: historically unequal access to healthcare (and historical bias) has resulted in a healthcare system in which much less money is spent on Black (vs. White) patients. Importantly, the researchers showed that lessening reliance on this proxy would increase the percentage of Black patients receiving additional help from 17.7 to 46.5 percent, virtually eliminating bias in the algorithmic outputs.[84]

Concerns with unintended bias extend to concerns with the use of technology in the criminal justice and legal realms. A few years ago, the Los Angeles Police Department implemented a new program to combat property crime.[85] The program relied upon an algorithm that had been trained on 10 years-worth of crime data and could produce a prediction as to where and when crimes were likely to occur. These areas of likely criminal behavior—"hotspots"—were subjected to increased police scrutiny. Protesters alleged that the software provided a patina of objectivity to what was mere bias: officers wanted to over-police predominantly Black neighborhoods, and the software gave them a seemingly objective reason for doing so.[86] In addition, the algorithm in question suffered from serious defects, as it relied on proxy data (a significant problem, as seen in the medical example in the previous paragraph) and focused on outcomes that were arguably the wrong ones on which to focus.

Bias in predictive analytics also has been alleged in pretrial risk assessment, systems designed to algorithmically generate "risk" scores indicating whether a defendant should go free or be detained while awaiting trial.[87] While there are many different risk assessment systems, all of them use some form of automation to generate outcomes, such as recidivism or dangerousness scores, that impact defendants' freedom and well-being. Such algorithms have faced near-constant scrutiny,[88] including allegations that they work to hide "overt discrimination based on demographics and socioeconomic status."[89] A 2016 report by *ProPublica* alleged that an algorithm used in Florida was biased against Black defendants,[90] such 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 were more likely than Black individuals to be labeled low-risk but then reoffend.

In 2019, prominent researchers signed a statement calling for a turn away from pretrial risk assessments and towards other reforms.[91] However, we conclude that, given the potential of AI and algorithmic decision making for improving decision making, retreat from these tools would be a mistake. As discussed in the remainder of this report, there are ways to design and implement them so their potential for good, rather than ill, is harnessed.

2
Guiding Principles for Using AI to Remove Bias in Decision Making

Given the overwhelming evidence, it is safe to assume that most government programs, when evaluated, will show signs of racial bias. Data and artificial intelligence can help government leaders correct for such biases. To do this, they need to employ AI or advanced computational and statistical techniques to train models that draw on an agency's own historical data. Just as hospitals and other organizations are using AI to regularize and optimize behavior in order to meet performance and profit goals, the same is possible when the goal is racial equity.

USING AI TO REDUCE VARIATIONS IN SURGICAL OUTCOMES

In 2012, a hospital network in Utah built an AI-backed system that provided recommendations for reducing variation in how surgeons performed certain medical procedures.[92] In essence, the machine generated anchors to guide and standardize surgeon behavior. Initially, the surgeons resisted the idea, believing that the machine would not be able to account for nuanced differences among patients and pathologies. However, this reluctance of experts to rely on something other than their own expertise was overcome, largely through the proven success of the intervention. The hospital system provided the surgeons with dashboards that relayed the statistical information, and the surgeons held weekly meetings to review the data and share the newly emerging best practices: that is, practices that improved patient outcomes while also decreasing costs. Over the past four years, the system has significantly improved the hospital's surgery results, and it has saved the hospital more than $90 million.

This initiative worked because: (1) it established consensus among doctors up front about what the optimal outcomes were. This promoted active anchoring on those optimal outcomes. In turn this (2) set a foundation for decision makers to overcome their own personal bias and more general human variance in decision making. This is precisely the route we propose for government organizations more broadly that are trying to overcome racial disparity in their employees' decision making and outcomes.

In the case of prosecutors in district attorney offices, there is often a lack of consensus about what optimal outcomes should be. For example, which criminal fact patterns deserve which outcomes? The answer is often beholden to various objectives and idiosyncrasies, such as office policy, an individual prosecutor's beliefs about the purpose of incarceration, a prosecutor's experience with such matters, his or her personal and professional experiences, and so on.

Whatever the optimal outcome might be for a fact pattern, it is possible for district attorney offices to ensure that similarly situated individuals are treated similarly. That is, the optimal approach can be defined as when differences in outcomes for similarly situated individuals are not driven by racial differences. These optimal anchors can be established using advanced analytics and AI. Then, when a new case arrives, a prosecutor can be shown multiple potential outcomes, all computed by the AI system, that show: (1) how the case would have resolved given the historical record, and (2) how the case would resolve if the defendant were treated according to the guiding principle of achieving racially fair outcomes (what can be called "counterfactual analysis"). These potential outcomes will then guide prosecutorial decision making, enabling them to overcome personal bias and more general human variance in their decision making.

The next section of this report describes a three-part framework for using data and AI to anchor decision makers on fair and just outcomes. However, this framework is premised on two principles that guide the design of all three parts of the framework: transparency and accountability. The first of these guiding principles is transparency.

Guiding Principle One: Transparency

A significant worry about artificial intelligence is that it operates in the dark, that it is "black box decision making."[93] Some AI processes, especially deep learning—that is, machine learning methods that are based on artificial neural networks—are indeed somewhat opaque.[94] This opacity, what can be called the problem of "explainability," is an important subset of transparency.

A 2019 law review article[95] discusses the need for two types of transparency:

- "Fishbowl Transparency" requires information about *what* officials are doing. The mandate for Fishbowl Transparency comes from a series of federal statutes, such as the Government in the Sunshine Act and the Federal Advisory Committee Act, and the Freedom of Information Act (FOIA). For this report, in which we zoom in on prosecutorial decision making, it must be noted that FOIA contains exemptions for law enforcement protocols.[96]
- "Reasoned Transparency" requires information about why officials took the actions they took. The mandate for Reasoned Transparency primarily comes from the due process clauses of the Fifth and Fourteenth Amendments to the U.S. Constitution.

When government decisions affect individuals' "life," "liberty," or "property" interests, a number of important legal protections apply:

- Substantive due process protections dictate that the government must show that its action was justifiable as a rational means of achieving a legitimate governmental purpose.[97]
- Procedural due process protections require the government to provide the affected individual with notice of its action, an opportunity to be heard, and procedural consistency, such that the individual received the same procedure as others and was not subject to a procedure designed to disadvantage him or her specifically.[98]
- Administrative/procedural protections are embedded in the Administrative Procedure Act (APA),[99] (however, these are not germane for prosecutorial decision making).[100]

In the case of algorithmic decision making, due process standards can be met when machine decisions are sufficiently transparent and explainable. When this occurs, the legality and finality of decisions can be defended by a showing that the process was valid.[101] In particular, a district attorney's office might think through the *Mathews v. Eldridge*[102] due process balancing test, in which three interests are balancing:

- the private interest at stake;
- the risk of error as a result of the administrative procedure and the likely error-correction benefit of a proposed change; and
- the government's interest.

When machine decisions are not sufficiently transparent and explainable, it will be very difficult to satisfy due process concerns,[103] not to mention the APA's procedural requirements.[104]

If Fishbowl Transparency requires disclosure of *what* officials are doing, at a minimum this would include a duty to inform the public that algorithmic decision making is being used. More than a bare acknowledgment of use, however, this should be genuine disclosure: agencies must publicize when and where AI systems are used—and for what purpose. Public discussion of the propriety of such use should be encouraged.

Reasoned Transparency, requiring disclosure of *why* officials did what they did, is more complex. At a minimum, it is certain that government technology cannot be entirely withheld from public inspection. For example, a Texas school district that used AI for teacher evaluations was forced to discontinue use of the AI when its developer refused to share the proprietary formula that generated the output.[105]

More importantly, violation of Reasoned Transparency raises deep questions relating to the imperative—long established in U.S. law—that government decision making be explainable.[106] Almost all technology used in the legal realm makes use of computer code. Some of this code, and some of the algorithms, are exceedingly complex and difficult to parse rationally. At the beginning of this section, we referred to this as the black box problem. For example, neural nets are powerful tools, but they are not designed to produce consistent or even replicable results. Thus, explaining complex computational models may seem like a Sisyphean task. In *State v. Loomis*,[107] Justice Shirley Abrahamson of the Wisconsin Supreme Court wrote in a concurrence, "[T]his court's lack of understanding of COMPAS [a software package] was a significant problem…The court needed all the help it could get." Because of accounts like this one, some have argued that district attorney offices should hesitate to use complex technologies.[108] This is an easy action point, and we understand the appeal of suggesting that offices need to use the most scrutable of algorithms that they can.

However, instead of demanding interpretability and then building a model, organizations such as district attorney offices should first build models and then work on interpretability—prior to using the model in actual cases and matters, of course. This is because there is an art to modeling. Eliminating racial bias and maximizing prediction is a difficult computational task. The first objective should be to achieve it. The second objective should be to unpack what the model is doing.

Returning to *State v. Loomis* and Justice Abrahamson's concurrence, what happened there was not an impossibility of explaining the software's results. Rather, it was a failure to explain. After all, the model actually was not complex and could have been explained, if only the right translator[109] were tasked with the job. Further, a team of computer scientists in 2019[110] have shown that, even with extremely complex models, there are methods of dividing the models into smaller subspaces which can be explained. In turn, these explanations allow for intelligible explanation of overall model workings. This is especially illuminating when one realizes that the model developed by these computer scientists was significantly more complex than the COMPAS tool at issue in *State v. Loomis*. Moreover, research on explainability is rapidly advancing[111]: for example, Google has released powerful tools for explaining AI,[112] and the Defense Advanced Research Projects Agency and the National Science Foundation have jointly emphasized funding research into explainable AI.[113]

Finally, transparency requires public disclosure. Government agencies that rely on technology to support decisions affecting individuals should make both data and code publicly available. Of course, the data should be partially scrubbed and redacted for privacy reasons, but as much as possible should be disclosed, and the same is true for code. We follow the lead of

others here in suggesting that offices make available as much of the data and code as possible available; if it is not released, there should be detailed statements of the enforcement priorities and practices that formed the basis for the code.[114] In this way, auditing and outside review will be a matter of record. This is not a mere utopian vision, either. Some district attorney 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.[115] There are a number of national organizations—the IJIS Institute, the FBI's Advisory Policy Board, and countless others—that can help facilitate such transparency.

RECOMMENDATION

Ensure decisions made using advanced analytics are transparent and explainable. Specifically, agency leaders should:

- Inform the public that algorithmic decision making is being used.
- Publicize when and for what purpose such systems are used.
- Disclose information about why officials took the actions they took.
- Make both data and code publicly available. The data may need to be partially scrubbed and redacted for privacy reasons, but as much as possible should be disclosed.

Once the computational models are built, but prior to actual use, the agency should ensure that the models and outcomes are sufficiently transparent and explainable to satisfy legal due process procedures.

Guiding Principle Two: Accountability

Accountability requires setting clear metrics for success, assessing progress on those metrics over time, and establishing a fair process for appeals. If the goal is, for example, to rid district attorney offices of racial bias in case outcomes, then the question motivating the metric is clear: are similarly situated defendants being treated similarly? While there are a few different ways of assessing this, a district attorney's office must decide which route to take (and using more than one is certainly appropriate), and it must produce at least quarterly reports, similar to a financial accounting, that describe how the office is doing in terms of these metrics.[116]

Some researchers have argued for the use of Algorithmic Impact Assessments (AIAs).[117] AIAs would require extensive review of technological impact, and this review would be independent of the office's stated goals. For example, if an office uses technology to achieve greater efficiency (and not necessarily racial equity), then fairness concerns would still be monitored via AIAs.[118]

While district attorneys seldom like appeals or disturbances of seemingly settled matters, there must be a mechanism in place for those adversely affected when technology is used to support decisions. Accountability demands a robust appeals process. However, this does not have to be an expansive right that goes beyond what is already in place. In short, appeals of technologically-informed decisions should be tailored to accord with whatever rights of appeal a district attorney's office already grants in response to human decisions. This is especially true

when decisions are fully automated. As mentioned above, some disability determinations have been delegated to Automated Decision Systems (ADS). In a recent case, individuals who were adversely affected by ADS decisions were not given insight into the machine's decision making and had no meaningful course of appeal.[119] Regardless of whether or not this particular appeal/remedy problem will be addressed legislatively,[120] any government organization that uses technology to make decisions that affect an individual's rights must be accountable, and accountability requires a right of appeal that tracks whatever right attaches to human decisions of a similar type.[121] At a minimum, defendants should be allowed to contest inaccuracies in machine inputs and to present mitigating information.

RECOMMENDATION

Ensure advanced analytics include strong accountability for their use. Specifically, agency leaders should:

- Set clear metrics for success and assess progress on those metrics in quarterly reports, similar to that done in financial accounting practices.
- Establish a right of appeal that tracks whatever right attaches to human decisions of a similar type. At a minimum, complainants should be allowed to contest inaccuracies in machine inputs and to present mitigating information.

3
Creating a Framework for Using AI to Remove Bias in Decision Making

With the two guiding principles of transparency and accountability as the context, the following three steps will result in a framework that will allow government organizations to use their own data to anchor decision making on fair outcomes.

- Step One is to collect the appropriate data.
- Step Two is to construct the computational models.
- Step Three is to manage the human-computer interaction in making decisions informed by AI.

Step One: Collect the Data

When it comes to processes that are internal to government, data collection is surprisingly uneven.[122] In the legal context, and especially in the context of criminal law, the availability of consistent data of sufficient quality is notoriously low.[123] Many offices, such as district attorney offices, keep scant data, and what data they do keep tend to be kept confidential. This has been the case for many years.

Data collection is, of course, necessary for interventions that rely on data, but mere collection 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 stand in the way of widespread adoption.[124] Given that machines learn from historical data, data that are unrepresentative threaten to introduce difficult-to-eradicate biases,[125] especially if the ways in which the data are unrepresentative are not known.[126]

What is ultimately needed? In short, widespread and standardized data collection that can be made available for public consumption. (For more on the types of data collection, see below, especially the discussion of efforts in Connecticut.) Widespread and standardized data collection is necessary because the effects of racial bias tend to only be observable at aggregate levels.[127] In order to evaluate whether implicit bias and group-level biases are affecting group-level behavior, large-scale multilevel data collection efforts are required.

This may seem like a daunting task, given that many government organizations use legacy systems and outdated IT infrastructures, and also given the relatively siloed nature of government data repositories. However, in the criminal justice domain and others,[128] it is not insurmountable. After all, courts—often through clerks of court—store case information, and most of it is stored electronically. In addition, given that these are criminal cases we are discussing in the context of prosecutorial decision making, the data is already largely in the public domain.

For insights into how these fragmented sources might be brought together, consider the healthcare system, where data are more voluminous and even more scattered than in the typical governmental realm. If a medical patient sees five different physicians, it is typical that the patient's data will be stored in five different systems.[129] Likewise, if an individual is arrested in different jurisdictions, his or her data will be stored by different clerks of court.

Following are examples of how disparate data sets from different systems have been integrated to improve decision making:

- Answer ALS, a global project focused on combating Amyotrophic Lateral Sclerosis (ALS), has developed a platform for bringing data from multiple centers (Cedars-Sinai Medical Center; the New York Genome Center; Massachusetts General Hospital) to a central repository from which it can be analyzed.[130]
- The U.S. Department of Transportation's National Address Database is another example of a collaborative, centralized, and readily available data effort, one that facilitated cross-agency development and was adopted by 22 states.[131]
- Even more impressively, the U.S. Agency for International Development has created an enterprise-wide data collection system that focuses on diverse topics and spans the globe.[132]

What these agencies are achieving in their own domains could be achieved in the legal world, as well.

DATA CENTRALIZATION CAN LEAD TO COST SAVINGS IN GOVERNMENT

Efforts to centralize and standardize data in order to facilitate the goal of reducing bias could have auxiliary benefits. A report by the Technology CEO Council found that such data efforts would lead to tremendous cost reductions in government: IT modernization would lead to an estimated $110 billion cost reduction over 10 years, shared services would lead to a $47 billion reduction, and analytics and cognitive computing would yield a $205 billion reduction.[133] This is unsurprising given that, in 2016, approximately 75 percent of government spending on IT was allocated to the operation and maintenance of ineffective and near-obsolete legacy systems.[134]

Even though much more needs to be done, significant progress in collecting standardized criminal justice data is underway. Following are a range of federal, state, and non-governmental examples to centralize and standardize the collection of these data:

- With support from the National Science Foundation, researchers at Northwestern University are compiling court data from a national repository and linking it to data about litigants, judges, lawyers, and courts.[135] The result will be an open, searchable, AI-powered data platform that will dramatically increase the accessibility and transparency of federal courts.
- The Bureau of Justice Statistics has instituted a data improvement program with the mission of improving the criminal record-keeping of states and local governments while improving the ability of states and localities to produce statistics on crime and the administration of justice.[136]
- The Connecticut legislature recently passed a law requiring prosecutors to compile data on how many defendants received prison time, plea bargains, or were diverted, and these datasets have to include information regarding defendants' race, ethnicity, sex, and age.[137] 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.[138]

- State representative Marsha Judkins from Provo, Utah, is working on a "prosecution transparency bill" that will require Utah counties to collect information on arrest, charge, sentence, and parole decisions.[139]
- The MacArthur Foundation is supporting Besiki Luka Kutateladze, a Florida International University professor, in an effort to improve data collection and analysis in four major prosecutor's offices: Jacksonville, Tampa, Milwaukee, and Chicago.[140]
- Measures for Justice, a data portal that is compiling court and corrections data at the county-level, has made significant progress in creating precisely what we envisioned: a national and centralized repository.[141]
- Likewise, Search, a nonprofit, is actively working with states to improve and standardize their criminal history record information.[142]

In addition to improving data quality, these large and centralized repositories can work to solve existing data quality issues. There are various data preprocessing approaches to fairness which can be used, given sufficiently large datasets, to identify and correct for unrepresentative or poor data.[143]

Even though centralized and standardized data collection is the ultimate goal, the immediate goal is to collect the data that are appropriate to the defined goals of the users. Just as the Connecticut legislature recognized, data must be collected by prosecutors' offices, but this collection does not have to be exhaustive. Only those variables that are essential for assessing bias are needed, including but not limited to initial charges and facts (such as arrest reports), case outcomes and method of resolution, as well as demographic information on defendants (including criminal history at time of arrest), prosecutors, judges, and arresting officers. Regardless of whether an office has historically been deficient at such data collection, proper collection should begin moving forward. The process does not need to be homegrown, either. For example, district attorney offices could reach out to nonprofits, such as Search, that are ready and willing to guide them.

RECOMMENDATION

Collect relevant data. Specifically, agency leaders should ensure:

- Initial data collection does not need to be exhaustive. The immediate goal is to collect the data that are essential to the purpose of the organization's AI initiative.
- The ultimate goal is to contribute to centralized and standardized data collection that is made available for public consumption.

Step Two: Construct the Models

Once an agency has completed the fundamental task of data collection, it then can turn to analysis. Analysis can be applied to any number of different objectives. In the case of a district attorney's office, it may want to analyze how much time its assistant district attorneys are spending on different types of cases. An office may want to use analytics to sort weaker from stronger cases, thus creating efficiencies in determining when to dismiss and when to pursue a matter further. In short, there are potentially many different goals when analyzing properly collected criminal justice data.

This report focuses on one particular aim: using a district attorney office's own data to correct for racial bias. For this task, there are two data analysis goals:

- First, identify whether the historical record suggests that there already is disparate treatment by race. In other words, is there evidence of racial bias? Within this task, the magnitude of the bias should be quantified. (There are countless methods for quantifying bias; as one example from the social sciences, see Kutateladze's work on prosecutorial bias.[144] See also the fairness research underway in computer science.[145])
- And second, devise a strategy for using that evidence to change human decision making in the future. In other words, how can our knowledge of past bias be used to make decision making more equitable and fair in the future?

To accomplish these goals, an important threshold question that must be answered is: who will perform the data analysis and/or construct the model? Even though some offices, such as the District Attorney of New York, are equipped with a data analysis unit, many government agencies, including most district attorneys' offices, are not so equipped,[146] and thus collaboration with a third-party will be necessary.

While there are a host of data science and predictive analytics companies working in this space, another untapped resource is academia. Research scholars are professionally attuned to the principles of transparency and accountability discussed above. In addition, there are many interdisciplinary academics working in this field. For example, in the case of district attorney offices, they could reach out to departments of criminology, psychology, law, applied mathematics, politics, economics, sociology, or other related fields. The prospect of having a unique dataset to work with may be more than enough to incentivize research scholars to participate.

In determining whom to work with, government leaders could contact their state, city, or agency's Chief Data Officer (CDO). These individuals have the authority and express mandate to advance data-driven government, and thus could function as valuable resources.[147] Numerous states, cities, and agencies have CDOs and have had them for a few years now.[148] New York City was the first city to create a CDO position, doing so in 2011, and since then CDOs have distinguished themselves for combining technical expertise and community engagement, enabling their communities to achieve their goals by harnessing data.[149]

Another threshold question concerns what the data analysis and modeling will consist of. Answering this question requires balancing input from two parties. In our example, on the one side, there are the academics and researchers, who are the technical experts in areas such as data science.[150] On the other side, there are prosecutors in district attorney offices, who are the subject matter experts. Understanding the role of these two parties will largely explain how modeling should proceed. In following two sections, we discuss these respective experts.

However, it is worth noting that, while we present a specific approach for meeting these goals, it is far from the only approach that can be taken. We encourage district attorney offices that seek to meet these goals to be open and to experiment with their academic research partners—a novel approach could very well turn out to be the best approach. That said, we are proposing that offices create, via computational modeling, a simulation for prosecutors, one that reveals how new cases would resolve if defendants were treated fairly in terms of race.

The Role of Technical Expertise.

Technical expertise is an essential element to designing any system relying on artificial intelligence to support decisions. In this case, we focus on the role of technical experts in designing a system to remove racial bias from decisions made by prosecuting district attorney offices. This involves several technical tasks:

- ***Identify and quantify existing bias.*** The first task of the technical experts is to determine whether there is evidence of pre-existing racial bias in a district attorney office's decision making. In the present example, this determination is complex given that there is a lack of a true baseline: one does not know who actually did or did not commit the alleged crimes. In addition, there are other decision makers involved upstream (namely, police officers and individuals working in pretrial services), and their bias or lack or bias complicates our understanding of subsequent prosecutorial decision making. That said, there are many ways of controlling for these factors and for evaluating whether prosecutors are treating similarly situated defendants similarly. There have been at least 34 empirical studies published between 1990 and 2011 that explored this very issue.[151] In short, such analyses have been conducted again and again, and most any collaborating research partner will be able to assess a district attorney office's data for racial bias.
- ***Define human-centered "soft goals" to be achieved***. The second task is to address the "human problems." District attorney offices must clarify their internal objectives, rules, and priorities. Should arrests for certain drug crimes be treated with leniency? Should defendants who have criminal histories that include recent violent felonies be less likely to receive charge reductions? A 2016 *Harvard Business Review* article discussed "soft goals," that is, goals that an organization has but which have not been rendered into specific objectives. One might think of these as operating in the background as assistant district attorneys make case decisions. These are goals that must be fleshed out; prosecutors must ask, answer, and codify these questions. Only then can machine learning be employed to scale and regularize them.[152] Defining soft goals is especially important because some of them may conflict with the goal of bias reduction. When this is the case, best practices require identifying this tension and unpacking it in light of the quarterly reports. In other words, the dissonance caused by the conflicting goals is brought to light, and the effect of the soft goal on the bias reduction goal is quantified.
- ***Develop a strategy to deploy AI on the job.*** The third task is to devise a strategy for using the data to change decision making moving forward. There are reasons to be cautious about employing technical solutions to debias algorithms,[153] and there definitely should be direct challenges to techniques developed for debiasing algorithms,[154] but these are reasons for being cautious and for acknowledging hurdles. They are not reasons to forgo the use of technology in a legal setting altogether. Indeed, much progress is being made.[155]

When the technology employed makes use of predictive analytics, as in the intervention we propose, there are numerous methods for addressing bias in the historical record.[156] When the truth of a matter is not known (for example, we do not know whether a defendant actually committed a crime; we only know about arrest and conviction rates), fairness might require altering the underlying data to prevent the classifier from relying on attributes, such as race, that are inappropriate.[157]

A 2013 research study framed the problem as an optimization problem: how to alter the data so as to obfuscate information about membership in a protected group while also encoding the data as faithfully as possible.[158] Two European scholars, Faisal Kamiran and Toon Calders, in a classic attempt from a decade ago, "massaged" their dataset by making the least intrusive modifications that would result in an unbiased dataset.[159] On this modified

dataset, they then trained their classifier, which subsequently was able to significantly reduce bias in future classification without losing much predictive accuracy.

Likewise, a 2019 applied statistical study[160] developed a procedure for removing all information about race from the covariates used for prediction, thus guaranteeing similar distributions of the outcome variable (in this case, estimated risk of re-arrest). The study's objective was to take a dataset and construct a new dataset that contains no information about race so that any predictive model satisfies demographic parity (i.e., the requirement that predictions be independent of race). The study adopted this definition of fairness because its model was predicting rearrest, and we had no good way of knowing rates of criminal offending by race (this information is simply not known). In essence, and without going into the mathematical minutiae, we adjusted each variable by matching its estimated quantile (which depended upon the protected variable as well as the previously adjusted variables) to the marginal quantiles for that variable. The result is quite remarkable: the predictive accuracy of the method decreased only slightly due to the adjustment, but the adjusted model produced nearly identical predictive distributions by race.

Adapting the Data. Of course, there may be other underlying problems with the data that warrant alteration. For example, proportional bias (what is sometimes called "small data" or "low data" bias) also should be taken into account. There may be an underrepresentation in a given data set of, say, Black female insider traders. This is problematic because predictive algorithms will give low aggregate weight to rare groups or classes, such as Black female insider traders in our hypothetical example and will perform worse on them. To deal with such disproportionate data, the data might be altered, or the model might be designed so as to correct for deficiencies in the data. For example, in 2019 a team of researchers created an algorithm that combines an original learning task with a variational autoencoder in order to learn the latent structure of the dataset.[161] 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.[162]

Adapting the Models. This leads to the second approach to fairness when the truth of the matter is not known: adapt the model so that similarly situated individuals are treated similarly.[163] Many of these approaches begin by 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—which might function as mere proxies for race). In 2011, two economists[164] developed a method in which they use information from illegitimate (or "contested") predictors, but they debias the predictors by marginalizing their importance in the model. A 2014 study[165] measured the effect of an illegitimate predictor in a subset of the dataset using an estimated probability distribution. Then they proposed a classification method that corrected for the discovered discrimination without using the protected attribute in the decision process. Similarly, a team of researchers in 2017[166] suggested that a weak version of disparate treatment access an illegitimate predictor during training but omit the attribute during classification. In other words, the illegitimate predictor (race and its covariates) is included when the model is being trained so that predictive accuracy is optimized. But then, when the model is used with new cases, race (and its covariates) is not input into the model.

As another promising example, a group of scholars in 2019 created a Bias-Resilient Neural Network (BR-Net), where problematic features, such as race, are identified by humans, and the BR-Net then learns to prevent improper correlations between those variables and the output.[167] This process works as follows. First, a model is trained to maximize prediction. Second, the model identifies the correlations between the bias variable and embeddings to other variables. Third, the model minimizes the influence of the bias variable by generating

new embeddings that do not correlate with the bias variable. In total, the BR-Net is able to generate embeddings that maximize model performance while minimizing the biased correlations. The researchers tested the BR-Net using photo gender prediction, as these classifiers usually perform much worse with Black individuals. The BR-Net was able to yield results that were 2 percent more accurate across race than a similar model that did not have the BR-Net addition.

Regardless of the approach taken, it is clear that fair classification entails trade-offs.[168] This is the case because there are a number of reasonable ways of evaluating a model for fairness, including ones that draw on calibration, false positive rates, false negative rates, and so on, and these desiderata typically cannot be achieved simultaneously.[169] For example, two Israeli computer scientists in 2018 focused on achieving similar false positive rates and similar false negative rates in different populations, an approach that theoretically makes sense in the criminal justice context,[170] but this does not mean that all parties will agree that such an algorithm is fair. It is this potential for disagreement brings us to our subject matter experts in the next section.

Finally, technical experts should be in charge of translating the models into software code and they will design the interface by which prosecutors will use the tools. Such interfaces raise issues of data security and data privacy, which will need to be handled in accordance with how the office employs its other data tools, such as the office's case management system. A district attorney's office may also seek guidance from various in-house government technical experts, such as 18F, a U.S. General Services Administration group that provides federal agencies with advice on just such matters: modernizing software development processes, improving public-facing services like websites, and digitizing and streamlining internal systems.[171]

The Role of Subject Matter Expertise.

The previous section on technical experts describes several debiasing methods that involve 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—which might function as proxies for race). While the example of zip code may make it seem as if it were simple to determine which predictors are legitimate and which are not, this task is actually quite difficult. Columbia University law professor Bernard Harcourt,[172] for instance, argued that criminal records are proxies for race. However, an argument can be made that criminal convictions are legitimate predictors but criminal *arrests* are not, since the latter may suggest police bias. What should be done? Questions like this one—how should criminal records be handled?—require significant subject matter knowledge in the area[173] and are best posed to subject matter experts. That is, prosecutors in district attorney offices and other key stakeholders, such as community members, law enforcement subject matter experts, and victim advocates brought into the discussion, must decide, and they must be willing to openly discuss the basis for their decisions.

Previously, we described a type of bias that emerges from training a model on data with base rates that are liable to lead to incorrect conclusions about causality. 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.[174] If the domain is law, and the model is evaluating drug offenses, then lawyers who understand historical racial bias and the fact that minorities—and especially Black individuals—have been over-policed and over-prosecuted for marijuana offenses, must be involved in critiquing the outputs.

No matter how well-collected the criminal justice dataset, any technology that builds upon it must be vetted by experts who can provide context. This is because, at the least, technology should operate fairly, and fairness must be defined by key stakeholders.[175] Once it is so

defined, then context becomes vital for making sure that the technology used for legal decision making employs the fairness definition that the key stakeholders think is the appropriate one to use.[176] 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 its implementation will require careful reworking of the algorithm.

Importantly, these discussions, which appear to be policy-focused, dovetail with the technical discussions raised in the previous section. For instance, computer scientists Sam Corbett-Davis and Sharad Goel have argued that anti-classification (when race is not explicitly used to make decisions), classification parity (when, for example, false positive and false negative rates are equal across groups), and calibration (when outcomes are independent of protected attributes) all are limited statistically.[177] According to these scholars, the best approach is often the one in which similarly risky people are treated similarly, based on the most statistically accurate estimates of risk that one can produce. While this argument is not as germane in our present example, where defendants' "riskiness" is not necessarily a key consideration in prosecutorial decision making, it shows just how multifaceted and particular these discussions are.

In addition, note that, in this report, we are focused on what government decision makers have done and will do. Specifically, we are focused on what legal actors—prosecutors, for instance—will do: will they offer a charge reduction? Will they offer a favorable plea bargain? Will they impose a prison sentence or a non-carceral sentence? These are the kinds of predictions from which bias can more easily be removed, and these are the ones that are most relevant to our report.

In contrast, risk assessment tools, which are often used by judges, represent a different form of prediction: now the prediction concerns what people beyond the decision makers will do. For example, what will defendants do: will they show up for court? Will they commit additional offenses if released? It's more difficult to eliminate racial bias in these kinds of predictions. To address such difficulties, organizations, such as the Partnership on AI, have created guidelines for using risk assessment tools.[178]

Granted, some of the former type of predictions (what decision makers will do) receive the latter (what others will do) as inputs, and this is something to which subject matter experts must be attuned and must identify for and discuss with the technical experts.

In all, technical experts and subject matter experts must work together to decide upon an approach that they think is best. This approach must then be opened to key stakeholders, including the public, so that, in turn, it might be further vetted and refined by the experts, such as through the use of agile methods. While we discuss this process in more detail in Step Three, we should mention that, even in the absence of machine decision making, these hard decisions are already being made. When humans make decisions, they are precipitated by underlying beliefs regarding various factors, including fairness. The primary difference between human decision making and the process we mention in here and in Step Three, is that the algorithmic process, with its emphasis on reaching a definition and instantiating it in code, is more transparent. With machines designed in the way we have outlined, we know what decisions have been made. In contrast, with humans acting alone, bias is shrouded, and motives and beliefs are obfuscated.

RECOMMENDATION

Build computational models that reflect key values, such as fair outcomes. Agency leaders need to articulate, up front, the values that their decision process should reflect. Specifically:

- When correcting for racial bias, the overarching objective should be to build a model that can anchor decision making on fair outcomes.
- Set clear objectives and, when predictive analytics are employed, clearly define what is being predicted.
- Form a team of technical experts to handle model construction by partnering with third party experts, such as academics.
- Form a separate team of subject matter experts to (1) explore policy questions that require domain knowledge, (2) interface with community members and key stakeholders in ensuring fairness, and (3) work closely with the technical experts to ensure that the models are built in accordance with their intended design.

Step Three: Manage the Human-Computer Interaction

Whenever people are expected to work closely with algorithmically-generated recommendations, there should be a "reasoned rule."[179] A reasoned rule is basically a written account of an office's agreed upon decision making protocol. Developing one involves selecting a few variables that are incontrovertibly related to an outcome. These variables should be assigned equal weight in the prediction formula, and their valence should be positive or negative depending on the specific variable.

For example, a history of violent felony convictions would be negative (less likely to result in a charge reduction), while a more severe charge of a drug crime would be positive (more likely to result in a charge reduction). An interdisciplinary team of researchers writing in the *Harvard Business Review* in 2017 emphasized that such rules are valuable because they, one, enable decisions that can be made quickly and without a computer; two, require only limited information to reach decisions; and, three, allow for insight into justifications for the decisions made.[180]

Reasoned rule models have proven successful in judicial decision making,[181] personnel selection, election forecasting, and many other tasks.[182] The collective use of the computational model (developed in Step Two, above) and a reasoned rule model is what would be called "ensemble modeling," where the goal is to engage multiple models that enable evaluation and comparison so that greater insight into decisions can be gained.

Interactive Training.

In what setting will the bulk of this evaluation and comparison take place? We recommend that it take place in a series of interactive training sessions. In these sessions, prosecutors could initially view hypothetical cases and compare their instinctive charging and plea bargaining decision making versus the reasoned rule and also the counterfactual outcomes produced by our computational model. This would have a number of benefits. It would encourage prosecutors to think deeply about cases and outcomes. It also would serve to vet and tweak the computational model. If an attorney's decision making deviates from the model's, the difference could be discussed in the learning environment. Whether the model

should be altered or whether the human should alter course would be up for debate. After many iterations of this process, accompanied by subsequent tweaking of the model when warranted, the model's robustness should be greatly improved.

Flexibility and informality are strengths of this process. The models we have described should not be viewed as finished products. There must always be experimentation and thoughtful iteration. For example, the New York City Fire Department uses an algorithmic system to predict outbreaks for fires in the city. While the initial algorithm was useful, the current iteration is 10 times more powerful, and there are plans to improve it even further.[183]

Subsequent training sessions should be divided into two groups: an experimental group, which would use the computational model to evaluate hypothetical cases, and a control group, which would not use the model. By tallying the performance of each group, this experimental process would enable vetting of collaborative intelligence. Using metrics the district attorney's office will have developed (see Recommendation One), the office must begin to answer the question of whether the algorithm is working as intended. The office should, at this point, begin to produce quarterly reports that reflect how the algorithm is influencing decision making. This is the case even though these reports will only reflect performance in training sessions and not actual performance. The point is that a district attorney's office should start producing these reports (revealing bias in its decisional outcomes) long before it incorporates the AI intervention into actual practice by a prosecuting attorney.

Even when a valid computational tool is developed, decision makers may not use it as intended. Two legal scholars in 2018[184] found that Virginia judges deviated from guidance provided by risk assessment tools. Likewise, a review of more than 1,500 bail decisions in Chicago found that, 85 percent of the time, judges did not follow a risk instrument's recommendation.[185] Since part of the motivation for adopting artificial intelligence is to decrease variation arising from human discretion,[186] we must be careful that, although use of such tools will be discretionary, prosecutors should be discouraged from consistently ignoring the system. In the experimental training sessions, this potential problem should be monitored and addressed. Indeed, the problem of noncompliance with the tool, if it arises, may have less to do with willful noncompliance than with discomfort with quantitative or statistical information.[187] For this reason, basic training should include information on how to interpret statistical confidence intervals, error rates, and the like. In the next section of this report, we provide an example to show how information might be provided so that it is not overwhelmingly quantitative/statistical.

Once this is accomplished, an organization is ready to begin to think about the human-computer interface; that is, it will be ready to integrate the system into a district attorney's office workflow so that it can be used to augment human decision making in actual cases. Here, the focus shifts. Instead of thinking about why and what to implement, now the office must think about how: the specific factors that determine how legal tech is brought into daily office practice.

Implementation.

Collecting the data and building the computer models are not sufficient; employees will need to change how they work in order to incorporate algorithmic-based decision support tools into their day-to-day routines. This means that an organization's leadership will need to reset the organization's culture and maintain a consistent tone in order for implementation to succeed. In the case of a district attorney's office, no longer is it about finding or not finding racial disparity in outcomes. It is now about finding ways to correct such disparities. This requires the district attorney, as the head of the office, to articulate new and different priorities:

- ***Focus on data-driven decision making.*** First, the district attorney must champion the need for change from the present focus on experience-driven decision making (e.g., the decision making of more senior employees is disproportionately valued) to a focus on data-driven decision making. Employees should, to the extent possible, augment their decision making processes with data-backed algorithmic recommendations, where the aim is to arrive at better answers than the human decision makers or the algorithmic ones could reach in isolation.[188] Voluminous research has found that professional judgment improves when it is coupled with data-based analyses and guidance,[189] and researchers have known this for decades: in 1986, Gottfredson and Gottfredson wrote that "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, and to find otherwise in criminal justice settings would be surprising (at best) and suspicious or very likely wrong (at worst)."
- ***Make a clear commitment to decrease racial bias.*** Second, the district attorney's leadership team will be responsible for articulating a clear commitment for the district attorney's office. A recent report found that, while pretrial risk assessments may lead to biased outcomes if used carelessly, targeted and goal-oriented use is generally successful, leading to reduced pretrial detention and less racial disparity.[190] For our topic of interest, the vision is that AI is going to be used to decrease racial bias in decision making. This vision—a district attorney's office without bias—must be presented to the office staff 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 non-White individuals was significant historically, and it is significant today. AI shows the best potential for solving this problem, and it will enhance attorneys' abilities, not replace them. 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.
- ***Develop AI standards, processes, and policies.*** Third, the district attorney's leadership team must develop AI standards, processes, and policies. This includes addressing data privacy and protection concerns, of course, but it also involves addressing theoretical queries. Above, we discussed that there are many definitions of fairness. We 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 and thoroughly briefed by the subject matter experts, it is the leadership team that must ultimately provide answers and open up the answers to broader public scrutiny. In addition, the leadership team must consult with academics and lawyers in order to develop policies regarding algorithmic use that implicates Equal Protection Clause or Title VII concerns.[191] After all, algorithms that take into account race might be accused of being, in themselves, discriminatory.[192] There are strategies that district attorney offices could adopt. For one, the model could be used in an advisory capacity only,[193] as some scholars have argued that when algorithmic output does not directly determine what government action is taken, various demands, including ones for transparency, are somewhat relaxed.[194]
- ***Use "translators" to bridge between the old and new ways of doing work***. Fourth, the district attorney's leadership team must identify and prioritize "translators."[195] These individuals, who may be drawn from either the internal staff or from the ranks of the collaborating academic researchers, must bridge the gap between the data scientists and the employees. Their objective is to explain the details of the technology and why it is producing the output that it is producing. In addition, they must synthesize feedback from employees and present it to the data scientists. Translators also will be the primary party tasked with exploring unintended consequences. Whenever AI systems are designed and implemented, they often have consequences on employees and third parties that were not foreseen. These must be documented and considered carefully.

- ***Be transparent about the metrics.*** Fifth, the district attorney's leadership team must prioritize the transparency of the metrics they developed earlier in the process, and they must disseminate quarterly bias reports, similar to the way financial reports are made available. The reports should outline how the office is performing on metrics of racial disparity in outcomes. These reports also will form a foundation for allowing regular audits of the organization's performance and its use of AI tools.
- ***Budget for both the "hard" and "soft" parts of the initiative.*** Lastly, as a practical matter, district attorney offices should budget as much for the implementation phase as they do for the technology development phase. Leadership must be proactive and decisive in reviewing the results of these reports. If there are signs that the AI tools are leading to unintended consequences, are being used inappropriately, or are failing to decrease disparities, action must be taken. This may include immediate cessation of use, followed by a period of discussion and review by the full team of technical and subject matter experts and leadership. In a recent survey of AI adoption by private companies, it was found that nearly 90 percent of the companies that had successfully scaled AI had spent greater than 50 percent of their analytics budget on implementation.[196] While securing funding is never frictionless, there are number of avenues by which offices might secure funding for such initiatives. Depending on the field, there likely are national organizations who would support such work. In addition, if there are academic partners, their universities might have grants available, and offices can collaborative with academics to jointly apply to governmental and private funding opportunities.

  The lesson of this is that merely figuring out data collection and model construction is not enough; implementation is at least as important, and at least as costly, as these more visible undertakings.

Throughout implementation, an organization's leadership team should be patient. Integrating transformative AI technology into an organization's day-to-day work processes is a long journey, and it is an especially long one for organizations, such as most government agencies that are not "born digital."[197]

RECOMMENDATION

Manage the human-computer interaction. Agency leaders need to:

- Draft a "reasoned rule," a written account of an agency's agreed upon decision making protocol, which can be compared against both human and machine decisions.
- Introduce the computational model during training sessions with hypothetical cases, so that both human and machine decisions can be vetted, critiqued, and modified. This is when a significant portion of model refinement should occur.
- An agency's leadership will need to work with staff to change their mindset about bias from a focus on finding or not finding racial disparities in outcomes, to a focus on finding ways to correct such disparities by:
  - changing the culture from a focus on experience-driven decision making to a focus on data-driven decision making.
  - providing a clear vision, one that seeks to unite both leadership and employees on a common and exciting goal.
  - identifying and prioritizing "translators," individuals who can bridge the gap between experts and non-expert employees.
  - disseminating quarterly bias assessment reports.
  - budgeting for both the "hard" and "soft" parts of the initiative.

4

Applying the Framework in a District Attorney's Office

To provide a proof of concept as to how the framework could be used to remove bias in making prosecutorial decisions, we 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 report.

Collecting the Data

These data included case information for felonies that were charged in the district attorney office's jurisdiction 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.).

We found that the dataset lacked information regarding the criminal history of individual defendants, 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 this district attorney's office were to go beyond our proof of concept, the 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 manually 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). 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; a different district attorney office in another jurisdiction with which we have collaborated included police reports as text files within the cases' datafiles.

For simplicity's sake, in this example, we decided to focus on just one crime category: retail theft. For the same reason, we focused on only the two most common racial groups listed for defendants: White and Black.

Constructing the Computational 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.

We chose charge reductions as our 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.[198]

In building a model to predict charge reductions, it is tempting to assume that one can simply 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 geog-

raphy, 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.[199]

For our proof of concept, we decided to present prosecutors with a snapshot of three different augmented decision support models. For a new case, we show how the case would resolve if the prosecutor used:

- a predictive model that reflects how the defendant historically would have been treated.
- a race-neutral predictive model, which reflects one definition of fairness, and
- a "suggestive" model that shows how the case would resolve if the defendant were treated as if he or she were White.

Results of the Predictive Model.

Let's start with the purely predictive model. To build this model, we prepared our dataset by identifying directly or indirectly predictive variables, removing missing values, and creating a training and test split to avoid overfitting. Next, we chose the XGBoost algorithm, given its useful combination of accuracy and deft handling of overfitting. 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 far as accuracy, the model yielded an AUC (area under curve) of .86, which is considered good.[200] 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 defendants were Black individuals, they received charge reductions 11.2 percent of the time. For White defendants, this was 18.3 percent. With our predictive model, the rates for Black individuals and White individuals, respectively, was 11.6 percent and 22.8 percent.

Results of the Race-Neutral Predictive Model.

Second, we created another predictive model, except that in this one we showed how a new case would resolve if a defendant were treated race-neutrally. In designing an algorithm, it is important to consider the context of what is being translated into code. Constitutional scholar Aziz Huq recommends using different algorithms for different crimes and depending upon the undergirding policy goal.[201] For our model, we used "demographic parity" (i.e., the requirement that predictions be independent of race) as our fairness definition because, when it comes to predicting charge reductions, we really are predicting something that is largely independent of counterfactuals. If someone is charged with 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 our context here is very different than that of risk assessments, where discrete outcomes (e.g., recidivism, court appearance) are considered.

Even though we are using a specific type of classification system and definition of fairness, we are not exhaustively relaying the details of our approach with this dataset, as our approach is not germane to the project we are outlining. There are a host of modeling techniques available, and fairness must be determined contextually and reviewed *ex post*.[202] The particular approach a district attorney's office takes is up to the office, and we do not mean to constrain or influence such choices here.

In brief, for our race-neutral model, we used an approach that penalizes unfairness in the model training process. Specifically, we applied a method designed by algorithmic scholars Yahav Bechavod and Katrina Ligett,[203] 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. As these scholars note, this approach can be used with different types of models, including support vector machines (SVM) and, that which we use here, logistic regression. We used logistic regression as our base algorithm because it was fully tested in Bechavod and Ligett's published work. This model is our modified predictive model: it describes, in essence, how the case would resolve if the defendant is treated race-neutrally.

We include this model because we believe it serves a cautionary tale. In terms of overall charge reduction rates, the model did not decrease racial disparity by much. 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.

Results of the Suggestive Model.

Third, and most importantly, we created a "suggestive model" for a new case. We show how that case would resolve if the defendant were White. To build this model, we 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.[204] In other words, we did not just change the defendant's race from Black to White; we also altered any covariates of Black race (e.g., incident location) so that they appeared to reflect White race. Importantly, we 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.). We only altered inessential features: ones that have no valid reason for influencing 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 jurisdiction and the judge attached to the case. After the transformation, a validation set of Black defendants was tested with the already trained model (the historical record predictive model, as described above).[205] Consider the impact of the transformation: converting the set of Black defendants to (seemingly) White defendants increased their probability of receiving a charge reduction by more than 5 percentage points, which represented a rate that was about equal to the rate for White defendants.

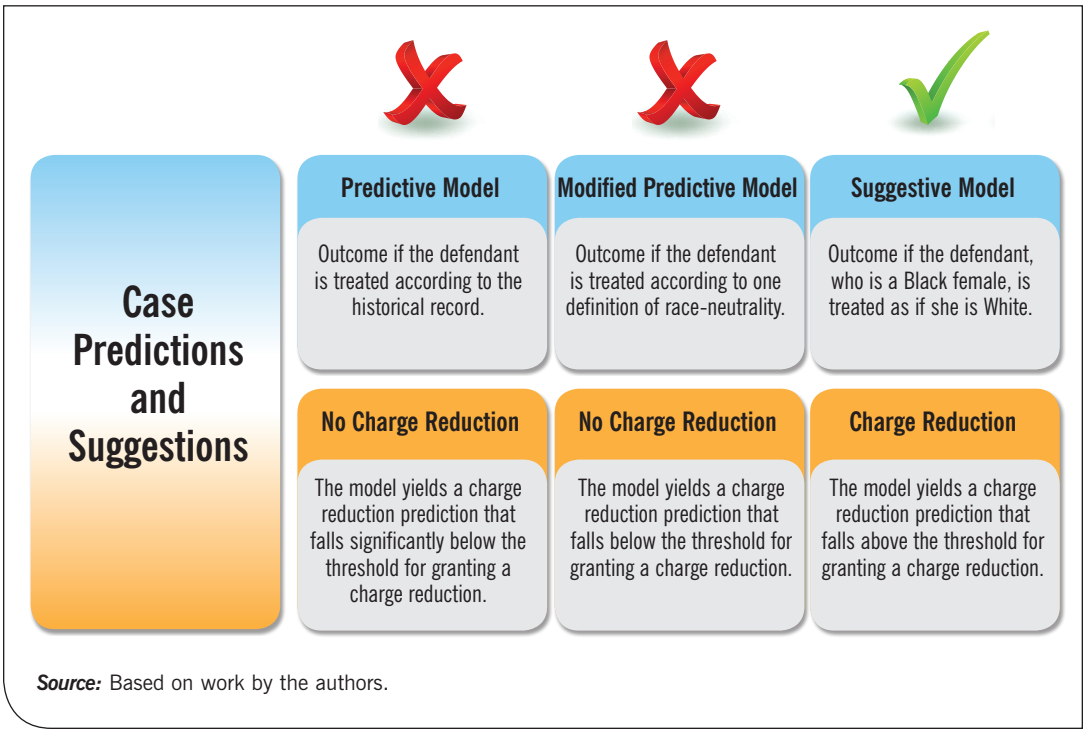
Addressing the Human-Computer Interaction

With these three computational models in place, we explored how they might be used for prosecutorial training and, later, actual cases. Consider the following case.

- On July 16, 2013, at 3 pm, 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 1: Case Predictions and Suggestions



Source: Based on work by the authors.

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, which fell below the machine-generated threshold for granting 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 we employed the procedure for transforming the defendant's race to White (and simultaneously changing the covariates so that they reflected, to the model, 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.

In addition, we don't belabor the factors and statistics because, in this report, we are approach agnostic, recommending neither this nor any specific fairness approach. Rather, in this report, we have outlined how an office might develop an approach that is tailored to its specific context and vetted, iteratively, in order to ensure fairness (see previous section, Creating a Framework for Using AI to Remove Bias in Decision Making). Indeed, we believe model creation is the core of the process and will require continuous discussion and revisiting.

With its model in place, a district attorney office may use the model to discuss decision making within prosecutorial training sessions. 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 augment prosecutorial decision making on new and current cases.

CONCLUSIONS

Government offices have access to stores of data—their own historical information—and this data should be viewed as an asset in the struggle for greater fairness and equality. When augmented with robust AI-equipped processes, such data can be used to create neutral suggestions to guide human decision making.

Such processes hold the potential for greater fairness and consistency in government decision making. But this is not a small task, and it is not one that can be undertaken without careful thinking. Unless AI systems are designed and implemented properly, they may perpetuate or even exacerbate the problems they were designed to solve.

This report presents two guiding principles for the design of an AI-assisted decision-making framework and three steps for building such a framework. The two guiding principles are transparency and accountability. The three steps for building a framework, which outline the specifics that should be considered are:

- collecting the data,
- constructing the computational models, and
- guiding the human-computer interface.

If government agencies follow this approach -- and especially if they do so in concert with academics, researchers, community members, key stakeholders, and other knowledgeable parties in government, such as agencies' Chief Data Officers—they will increase the chance of success in decreasing disparities in their agency's decisions and outcomes.

But this is just a starting point. Once data collection processes are more robust, government agencies can begin to assess more precisely how their decision making can be improved. Once the modeling processes are in place, agencies can deepen their engagement with community members and other stakeholders to determine the precise contours of what shape an AI initiative will take. And once the human-computer interface is ingrained in an agency's culture and workflow, then agency leaders can begin to think about future iterations. They can consult with Chief Data Officers and begin to formulate additional applications that will improve decision making. In other words, while this report represents merely a starting point, this start will enable creative, nimble, and fair processes in the near future.

ACKNOWLEDGMENTS

The authors are grateful to all those who took time to provide comments, critiques, and suggestions on ideas relating to this research, including J. Nicole Shelton, Stacey Sinclair, members of Princeton's Stigma and Social Perception Lab, members of the Princeton Project in Computational Law, and members of Princeton's Fellowship of Woodrow Wilson Scholars. In addition, the authors are grateful for funding to support the research that informs this report, including grants from Princeton's Department of African American Studies and the American Psychology-Law Society. Lastly, the authors would like to thank the Cornell Journal of Law & Public Policy for their input on ideas foundational to those presented in this report.

FOOTNOTES

1. Vijayan, J. (September 5, 2019), "Four Ways AI Can Augment Human Capabilities," *Information Week.* Retrieved from https://www.informationweek.com/big-data/ai-machine-learning/four-ways-ai-can-augment-human-capabilities/d/d-id/1335668.
2. Executive Order 13859, Maintaining American Leadership in Artificial Intelligence, Federal Register (February 14, 2019), v.84 no.31, p.3967, https://www.govinfo.gov/content/pkg/DCPD-201900073/html/DCPD-201900073.htm.
3. *E.g., see* 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/2018/10/22/business/efforts-to-acknowledge-the-risks-of-new-ai-technology.html.
4. 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.
5. 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.
6. Plenke, M. (September 9, 2015). The reason this "racist soap dispenser" doesn't work on Black skin. *Mic*. Retrieved from https://mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-Black-skin#.QU66mhRXv.
7. Plenke, M. (September 9, 2015). The reason this "racist soap dispenser" doesn't work on Black skin. *Mic*. Retrieved from https://mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-Black-skin#.QU66mhRXv.
8. 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, p. 133-134.
9. 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.
10. New York City Automated Decision Systems (ADS) Task Force. (November 2019). Report. Retrieved from https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf.
11. Sunstein, Cass R. (2019), Chapter 4,"Nudging: A Very Short Guide," *How Change Happens,* Cambridge: The MIT Press.
12. Labaton, S. (May 11, 1992). Benefits are refused more often to disabled Blacks, study finds. *The New York Times*. Retrieved from https://www.nytimes.com/.
13. Koch, J. W. (2018). Racial minorities' trust in government and government decisionmakers. *Social Science Quarterly, 100*(1), 19-37. doi: 10.1111/ssqu.12548.
14. 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.
15. Burke, K., & Leben, S. (2007). Procedural fairness: A key ingredient in public satisfaction. *Court Review, 44*, 4-25; Mitchell, O. (2005). A Meta-Analysis of Race and Sentencing Research: Explaining the Inconsistencies. *Journal of Quantitative Criminology*, *21*(4), 439-466. doi:10.1007/s10940-005-7362-7; Nellis, A. (2016). The color of justice: Racial and ethnic disparity in state prisons. *The Sentencing Project*. Retrieved from https://www.sentencingproject.org; The Sentencing Project. (2013). Report of the Sentencing Project to the United Nations Human Rights Committee regarding racial disparities in the United States criminal justice system. Retrieved from https://www.sentencingproject.org; Tonry, M., & Melewski, M. (2008). The malign effects of drug and crime control policies on Black Americans. *Crime & Justice, 37*, 1-44; United States Sentencing Commission. (2017). Demographic differences in sentencing: An update to the 2012 Booker Report. Retrieved from http://www.ussc.gov; Wagner, P. (2012, August 28). Incarceration is not an equal opportunity punishment. *Prison Policy Initiative*. Retrieved from www.prisonpolicy.org
16. United States Department of Commerce, Census Bureau. (2015). Quick facts: United States. Retrieved from https://www.census.gov/.
17. Carson, E. A. (2015). Prisoners in 2014. *Bureau of Justice Statistics*. Retrieved from https://www.bjs.gov.

18. Wagner, P. (2012, August 28). Incarceration is not an equal opportunity punishment. Prison Policy Initiative. Retrieved from https://www.prisonpolicy.org/articles/notequal.html.
19. Cohen, T. H., & Kyckelhahn, T. (2010). Felony defendant in large urban counties. *Bureau of Justice Statistics, US Department of Justice*. Retrieved from https://www.bjs.gov/content/pub/pdf/fdluc06.pdf.
20. Gramlich, J. (2019, June 11). Only 2 percent of federal criminal defendants go to trial, and most who do are found guilty. *Pew Research Center*. Retrieved from https://www.pewresearch.org/fact-tank/2019/06/11/only-2-of-federal-criminal-defendants-go-to-trial-and-most-who-do-are-found-guilty/.
21. Gilbert, S. A., & Johnson, M. T. (1996). The federal judicial center's 1996 survey of guideline experience. *Federal Sentencing Reporter, 9*, 87-93; 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; Miller, M. L. (2004). Domination and dissatisfaction: Prosecutors as sentencers. *Stanford Law Review, 56*, 1211-1276.
22. Albonetti, C. S. (1990). Race and the probability of pleading guilty. *Journal of Quantitative Criminology, 6*, 315–334; Albonetti, C. A. (1992). Charge reduction: An analysis of prosecutorial discretion in burglary and robbery cases. *Journal of Quantitative Criminology, 8*, 317–333; 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; 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; 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; 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.
23. Albonetti, C. S. (1990). Race and the probability of pleading guilty. *Journal of Quantitative Criminology, 6*, 315–334; Kellough, G., & Wortley, S. (2002). Remand for plea: Bail decisions and plea bargaining as commensurate decisions. *British Journal of Criminology, 42*, 186-210; 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.
24. Avery, J., & Cooper, J. (2020). Racial bias in post-arrest and pretrial decision making: The problem and a solution. *The Cornell Journal of Law & Public Policy.*
25. Allport, G. W. (1954). *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley; Williams, R. M., Jr. (1947). *The reduction of intergroup tensions*. New York: Social Science Research Council. *See also* Dovidio, J. F., Gaertner, S. L., & Kawakami, K. (2003). The Contact Hypothesis: The past, present, and the future. *Group Processes and Intergroup Relations, 6*, 5–21; Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65–85.
26. *See* Allport, G. W. (1954). *The Nature of Prejudice*. Cambridge, MA: Addison-Wesley; Cook, S. W. (1971). *The effect of unintended interracial contact upon racial interaction and attitude change* (Final report, Project No. 5-1320). Washington, DC: U.S. Department of Health, Education and Welfare, Office of Education.
27. Paluck, E. L., & Green, D. P. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology, 60*(1), 339-367. doi:10.1146/annurev.psych.60.110707.163607
28. Paluck, E. L., & Green, D. P. (2009). Prejudice Reduction: What Works? A Review and Assessment of Research and Practice. *Annual Review of Psychology, 60*(1), 339-367. doi:10.1146/annurev.psych.60.110707.163607
29. *See, e.g.*, Gardiner, G. S. (1972). Complexity training and prejudice reduction. *Journal of Applied Social Psychology*, *2*(4), 326-342; Gawronski, B., Deutsch, R., Mbirkou, S., Seibt, B., & Strack, F. (2008). When "just say no" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology, 44*, 370–377; Olson, M. A., & Fazio, R. H. (2006). Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin*, *32*(4), 421-433; Skorinko, J. & Sinclair, S. (2013). Perspective taking can increase stereotyping: The role of apparent stereotype confirmation. *Journal of Experimental Social Psychology, 49*(1), 10-18.
30. 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*. Advance online publication. doi: 10.1037/pspa0000160

31. 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*. Advance online publication. doi: 10.1037/pspa0000160
32. 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; 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; Lai, C. K., Hoffman, K. M., & Nosek, B. A. (2013). Reducing implicit prejudice. *Social and Personality Psychology Compass*, *7*, 315–330; 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; Sawyer, J., & Gampa, A. (2018). Implicit and explicit racial attitudes changed during black lives matter. *Personality and Social Psychology Bulletin, 44*(7), 1039-1059.
33. Lai, C. K., Skinner, A. L., Cooley, E., Murrar, S., Brauer, M., Devos, T., & Nosek, B. A. (2016).
34. Reducing implicit racial preferences: II. Intervention effectiveness across time. *Journal of Experimental Psychology: General*, *145*, 1001–1016. Smith, R. J., & Levinson, J. D. (2012). The impact of implicit racial bias on the exercise of
35. prosecutorial discretion. *Seattle University Law Review, 35*, 795-826. *See, e.g.*, 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; Eberhardt, J. (2019). *Biased: Uncovering the hidden prejudice that shapes what we see, think, and do*. New York: Viking.
36. 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
37. McKinley, J. C. (2014, July 8). Study finds racial disparity in criminal prosecutions. *The New York Times*. Retrieved from https://www.nytimes.com; *The New York Times* Editorial Board. (2014, July 14). How race skews prosecutions. *The New York Times*. Retrieved from: https://www.nytimes.com.
38. Bennett, M. W. (2010). Unraveling the Gordian Knot of implicit bias in jury selection: The problems of judge-dominated voir dire, the failed promise of *Batson*, and proposed solutions. *Harvard Law & Policy Review, 4*(1), 149-172.
39. Hetey, R. C., & Eberhardt, J. L. (2014). Racial disparities in incarceration increase acceptance of punitive policies. *Psychological Science, 25*, 1949–1954.
40. Hetey, R. C., & Eberhardt, J. L. (2014). Racial disparities in incarceration increase acceptance of punitive policies. *Psychological Science, 25*, 1949–1954.
41. Peffley, M., & Hurwitz, J. (2007). Persuasion and resistance: Race and the death penalty in America. *American Journal of Political Science, 51*, 996–1012.
42. *See* 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.
43. Gecker, J. (June 12, 2019). San Francisco prosecutors turn to AI to reduce racial bias. *Associated Press*. Retrieved from https://www.apnews.com/7e9305d3bccf4f8caee9960e772079f7.
44. 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.
45. Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. New York: Cambridge University Press.
46. 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; Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology, 3*, 430-454. doi:10.1017/cbo9780511809477.004
47. Kinder, D., & Sears, D. (1981). Prejudice and politics: Symbolic racism versus racial threats to the good life. *Journal of Personality and Social Psychology, 40*, 414-431. doi:10.1037/0022-3514.40.3.414.
48. 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
49. Simon, P. (March 2014). Potholes and big data: Crowdsourcing our way to better government. *Wired*. Retrieved from https://www.wired.com/insights/2014/03/potholes-big-data-crowd sourcing-way-better-government/.

50. GAO. (2004). Data mining: Federal efforts cover a wide range of uses. GAO-04-548, *U.S. General Accounting Office*. Retrieved from https://www.gao.gov/new.items/d04548.pdf.
51. The Department of the Treasury. (2009). Federal agency data mining report, 2008. *The Department of the Treasury*. Retrieved from https://www.treasury.gov/privacy/annual-reports/Documents/FY2008/DataMiningReport.pdf.
52. Kleinberg, J. Ludwig, J., Mullainathan, S., & Obermeyer, Z. (2015). Prediction policy problems. *The American Economic Review, 105*(5), 491-495; 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; Ravindranath, M. (September 28, 2014). In Chicago, food inspectors are guided by big data. *The Washington Post*. Retrieved from https://www.washingtonpost.com/.
53. Godtland, E., Grgich, M., Petersen, C. D., Sloane, D. M., & Walker, A. T. (2007). Racial disparities in federal disability benefits. *Contemporary Economic Policy, 25*, 27-45. doi: 10.1111/j.1465-7287.2006.00031.x
54. Berryhill, N. A. (2018). Social security administration: General statement. Retrieved from https://www.ssa.gov/budget/FY19Files/2019OP.pdf; See also Coglianese, C. (2019). Conference on social protection by artificial intelligence: Decoding human rights in a digital age. *Freedom to Tinker*. Retrieved from https://freedom-to-tinker.com/2019/05/29/conference-on-social-protection-by-artificial-intelligence-decoding-human-rights-in-a-digital-age/.
55. Bauguess, S. W. (2017). The role of big data, machine learning, and AI in assessing risks: A regulatory perspective. *U.S. Securities and Exchange Commission*. Retrieved from https://www.sec.gov/news/speech/bauguess-big-data-ai.
56. Berk, R. (2008). Forecasting methods in crime and justice. *Annual Review of Law and Social Science, 4*, 219-238.
57. Mayson, S. G. (2017). Dangerous defendants. *Yale Law Journal, 127*, 490-568.
58. Monahan, J., & Skeem, J. L. (2016). Risk assessment in criminal sentencing. *Annual Review of Clinical Psychology, 12*, 489-513; Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*, 803-872.
59. Klingele, C. (2015). The promises and perils of evidence-based corrections. *Notre Dame Law Review, 91*, 537-584.
60. Klingele, C. (2015). The promises and perils of evidence-based corrections. *Notre Dame Law Review, 91*, 537-584.
61. Ready, F. (April 10, 2019). California push to automate cannabis clearances portends future of conviction relief. *LegalTech News*. Retrieved from https://www.law.com/legaltechnews/2019/04/10/california-push-to-automate-cannabis-clearances-portends-future-of-conviction-relief/.
62. Ready, F. (April 10, 2019). California push to automate cannabis clearances portends future of conviction relief. *LegalTech News*. Retrieved from https://www.law.com/legaltechnews/2019/04/10/california-push-to-automate-cannabis-clearances-portends-future-of-conviction-relief/.
63. "Approximately 53,000 individuals will receive conviction relief through this partnership. Of those, approximately 32% are Black or African American, 20% are White, 45% are Latinx, and 3% are other or unknown." *See* Code for America's February 13, 2020 press release. Retrieved from https://www.codeforamerica.org/news/los-angeles-county-da-code-for-america-announce-dismissals-of-66-000-marijuana-convictions-marking-completion-of-five-county-clear-my-record-pilot.
64. Salomon, E., Bauman, M. J., Lin, T.-Y., Boxer, K., Naveed, H., … Ghani, R. (2017). Reducing incarceration through prioritized interventions. *Proceedings of ACM SIGKDD*.
65. Gecker, J. (June 12, 2019). San Francisco prosecutors turn to AI to reduce racial bias. *The Washington Post*. Retrieved from https://www.washingtonpost.com.
66. Arnold Foundation. (2018). *Public safety assessment: Risk factors and formula*. Retrieved from http://www.arnoldfoundation.org.
67. 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
68. Benjamin, R. (2019). *Race after technology: Abolitionist tools for the New Jim Code*. Medford, MA: Polity Press; David Freeman Engstrom, D. F., Ho, D. E., Sharkey, C. M., & Cuéllar, M.-F. (2020). Government by algorithm: Artificial intelligence in federal administrative agencies. Retrieved from https://www-cdn.law.stanford.edu/wp-content/uploads/2020/02/ACUS-AI-Report.pdf.

69. Executive Office of the President. (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. *The White House*. Retrieved from Retrieved from https://obamaWhitehouse.archives.gov/.
70. Ramirez, E., Brill, J., Ohlhausen, M. K., & McSweeney, T. (2016). Big data: A tool for inclusion or exclusion? Understanding the issues. Federal Trade Commission. Retrieved from https://www.ftc.gov/system/files/documents/reports/big-data-tool-inclusion-or-exclusion-understanding-issues/160106big-data-rpt.pdf.
71. Louridas, P., & Ebert, C. (2016). Machine learning. *IEEE Software, 33*(5), 110-115.
72. 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.
73. Dastin, D. (October 9, 2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*. Retrieved from https://www.reuters.com/.
74. Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments. *Sociological Methods & Research*. doi:10.1177/0049124118782533
75. 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.
76. 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
77. 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.
78. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*(3), 671-732.
79. Plenke, M. (September 9, 2015). The reason this "racist soap dispenser" doesn't work on black skin. *Mic*. Retrieved from https://mic.com/articles/124899/the-reason-this-racist-soap-dispenser-doesn-t-work-on-black-skin#.QU66mhRXv.
80. Garvie, C., & Frankle, J. (April 7, 2016). Facial-recognition software might have a racial bias problem. *The Atlantic*. Retrieved from https://www.theatlantic.com.
81. Vincent, J. (Jan. 25, 2019). Gender and racial bias found in Amazon's facial recognition technology (again). *The Verge*. Retrieved from https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender.
82. Grother, P., Ngan, M., Hanaoka, K. (2019). Face recognition vendor test (FRVT) Part 3: Demographic effects. *National Institute of Standards and Technology*; NLC Staff. (July 3, 2019). Racial bias in facial recognition technology: What city leaders should know. *National League of Cities*. Retrieved from https://citiesspeak.org/2019/07/03/racial-bias-in-facial-recognition-technology-what-city-leaders-should-know/.
83. 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
84. 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
85. Rieland, R. (March 5, 2018). Artificial intelligence is now used to predict crime. But is it biased? *Smithsonian*. Retrieved from https://www.smithsonianmag.com.
86. LA Times Editorial Board. (March 16, 2019). Editorial: The problem with LAPD's predictive policing. *Los Angeles Times*. Retrieved from https://www.latimes.com/opinion/editorials/la-ed-lapd-predictive-policing-20190316-story.html.
87. Hanna, M. (Feb. 6, 2018). What happened when New Jersey stopped relying on cash bail? *The Philadelphia Inquirer*. Retrieved from http://www2.philly.com/philly/news/new_jersey/new-jersey-cash-bail-risk-assessment-20180216.html; O'Dea, C. (July 31, 2018). Civil rights coalition calls for end to core element of NJ bail reform. *NJ Spotlight*. Retrieved from http://www.njspotlight.com/stories/18/07/30/civil-rights-groups-call-for-end-to-core-element-of-nj-bail-reform/.
88. 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; Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review, 104*, 671-732.

89. Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*, 803-872.
90. 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.
91. Minow, M., Zittrain, J., & Bowers, J. (2019). Technical flaws of pretrial risk assessments raise grave concerns. *Berkman Klein Center for Internet & Society at Harvard University*. Retrieved from https://cyber.harvard.edu/story/2019-07/technical-flaws-pretrial-risk-assessments-raise-grave-concerns.
92. Kass, E. M. (Nov. 18, 2019). Hospital system uses AI to boost surgery outcomes, cut costs. *The Wall Street Journal*. Retrieved from https://www.wsj.com/articles/hospital-system-uses-ai-to-boost-surgery-outcomes-cut-costs-11574073002.
93. Benjamin, R. (2019). *Race after technology*. Medford, MA: Polity Press.
94. *See* Mayer-Shönberger, V., & Cukier, K. (2013). *Big data: A revolution that will transform how we live, work, and think*. Eamon Dolan/Mariner Books; Pasquale, F. (2015). *The Black box society: The secret algorithms that control money and information*. Boston, MA: Harvard University Press.
95. Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review, 71*, 1-56.
96. Freedom of Information Act, 5 U.S.C. § 552(b)(1)–(9), 2018.
97. Mashaw, J. L. (2001). Small things like reasons are put in a jar: Reason and legitimacy in the administrative state. *Fordham Law Review*, 70, 17-35.
98. Shapiro, S. A., & Levy, R. E. (1987). Heightened scrutiny of the fourth branch: Separation of powers and the requirement of adequate reasons for agency decisions. *Duke Law Journal*, 387-455.
99. Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review, 71*, 1-56.
100. While the transparency demands outlined in this section may not be fully applicable in the law enforcement realm, criminal justice actors should strive to come as close as possible to satisfying them.
101. Morse, S. C. (2019). When robots make legal mistakes. *Oklahoma Law Review, 72*, 213-230.
102. *Mathews v. Eldridge*, 424 U.S. 319 (1976).
103. Morse, S. C. (2019). When robots make legal mistakes. *Oklahoma Law Review, 72*, 213-230.
104. *See* 5 U.S.C. § 706(2)(A). *See also* Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review, 71*, 1-56.
105. Webb, S., & Harden, J. D. (October 12, 2017). Houston ISD settles with union over controversial teacher evaluations. *Houston Chronicle*. Retrieved from http://bit.ly/2Iq3Ccx; *Houston Federation of Teachers, Local 2415 v. Houston Independent School District*, 251 F. Supp. 3d 1168 (S.D. Tex. 2017).
106. *See* 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; Selbst, A. D., & Powles, J. (2017). Meaningful information and the right to explanation. *International Data Privacy Law, 7*(4), 233-242.
107. *State v. Loomis*, 881 N.W.2d 749, 774 (Wis. 2016).
108. *See* Citron, D. K., & Pasquale, F. (2014). The scored society: Due Process for automated predictions. *Washington Law Review, 89*, 1-33; 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.
109. For more on translators, see Step 3: Manage the Human-Computer Interface, *infra*.
110. 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/~himalv/customizable.pdf.
111. Lipton, S. (2016). The mythos of model interpretability. *Proceedings of ICML 2016*. Retrieved from https://arxiv.org/pdf/1606.03490.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; Rudin, C. (2014). Algorithms for interpretable machine learning. *20th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 1519.
112. See https://cloud.google.com/explainable-ai/.

113. Turek, M. (2019). Explainable artificial intelligence. *Defense Advanced Research Projects Agency*. Retrieved from https://www.darpa.mil/program/explainable-artificial-intelligence.
114. Kleinfeld, J., et al. (2017). White paper of democratic criminal justice. *Northwestern University Law Review, 111*, 1693-1705.
115. See DA Stat. (2019). DA Stat: Data driven decision-making. *San Francisco District Attorney*. Retrieved from https://sfdistrictattorney.org/DAStat; Cook County Government. (2019). Open data. *Cook County State's Attorney*. Retrieved from https://datacatalog.cookcountyil.gov/.
116. This is discussed in greater detail in Step 3: Manage Human-Computer Interface, *infra*.
117. Reisman, D., Schultz, J., Crawford, K., Whittaker, M. (2018). Algorithmic Impact Assessments: A practical framework for public agency accountability. *AI Now Institute*. Retrieved from https://ainowinstitute.org/aiareport2018.html.
118. *See* Selbst, A. D. (2017). Disparate impact in big data policing. *Georgia Law Review, 52*, 109-195. doi:10.2139/ssrn.2819182
119. Lecher, C. (March 21, 2018). What happens when an algorithm cuts your health care? *The Verge*. Retrieved from https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.
120. Bowles, J. (April 14, 2019). Algorithmic accountability act targets bias in AI decision-making. *Diginomica*. Retrieved from https://diginomica.com/algorithmic-accountability-act-targets-bias-in-ai-decision-making.
121. Rights of appeal are an important and complex topic. One of the authors of this report is currently working on an article that more thoroughly unpacks this issue in the context of machine decision making.
122. *See* https://nces.ed.gov/partners/fedstat.asp for a listing of current resources.
123. 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/.
124. Olsen, R., Courtney, L., Warnberg, C., Samuels, J. (2018). Collecting and using data for prosecutorial decision-making: Findings from 2018 national survey of state prosecutors' offices. *Urban Institute*. Retrieved from https://www.urban.org/research/publication/collecting-and-using-data-prosecutorial-decisionmaking.
125. IBM Center for the Business of Government. (June 2019). *More than meets AI: Assessing the impact of artificial intelligence on the work of government, Part II.* IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/more-meets-ai-part-ii.
126. *See* Hao, K. (Jan. 21, 2019). AI is sending people to jail—and getting it wrong. *MIT Technology Review*. Retrieved from https://www.technologyreview.com/s/612775/algorithms-criminal-justice-ai.
127. 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.
128. For an example from a different domain, consider the Department of Defense, which has successfully integrated various data sources in a cloud-based, open-source platform in order to enhance its Biosurveillance Ecosystem. *See* DRTA. (July 6, 2017). Biosurveillance Ecosystem Enhancement - HDTRA1-17-RFI-CBI-BSVE - Federal Business Opportunities: Opportunities. Retrieved from https://beta.sam.gov/.
129. Stanley, A. (Jan. 9, 2018). Big Pharma seeks DLT solution for drug costs. *Coindesk*. Retrieved from https://www.coindesk.com/blockchain-day-big-pharma-seeks-dlt-solution-drug-costs/.
130. Others in the medical realm have recommended instantiating similar endeavors on a blockchain-based system, which would ensure data integrity. *See* De, N. (Oct. 24, 2017). HHS architect talks blockchain's potential role in healthcare administration. *Coindesk*. Retrieved from https://www.coindesk.com/hhs-it-architect-talks-blockchain-White-paper-results.
131. Wiseman, J. M. (2018). *Data-driven government: The role of chief data officers.* IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/data-driven-government-role-chief-data-officers.
132. Wiseman, J. M. (2018). *Data-driven government: The role of chief data officers*. IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/data-driven-government-role-chief-data-officers.
133. Technology CEO Council. (2017). *The government we need: How proven technology solutions can save taxpayers more than $1 trillion over a decade while enabling more effective government*. Technology CEO Council. Retrieved from http://www.techceocouncil.org/.

134. IBM Center for the Business of Government. (2017). *Transforming government through technology.* IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/transforming-government-through-technology.
135. *See* https://noacri.org/.
136. Bureau of Justice Statistics. Criminal justice data improvement program. Retrieved from https://www.bjs.gov/index.cfm?ty=tp&tid=4.
137. Mystal, E. (June 10, 2019). Prosecutorial data collection is coming to Connecticut. *Above the Law*. Retrieved from https://abovethelaw.com/2019/06/prosecutorial-data-collection-is-coming-to-connecticut/.
138. 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/news/2018/prosecutorial-attitudes-perspectives-and-priorities-insights-from-the-inside/report-1.pdf.
139. Mark Shenefelt, M. (2019, October 28). Prosecutor transparency bill planned for 2020 Utah Legislature. *The Daily Herald*. Retrieved from https://www.heraldextra.com/news/local/govt-and-politics/legislature/prosecutor-transparency-bill-planned-for-utah-legislature/article_8aba8e39-f8f8-5b77-a9da-64f7856ec60b.html.
140. Pantazi, A. (2018, March 9). What makes a good prosecutor? A new study of Melissa Nelson's office hopes to find out. *The Florida Times-Union*. Retrieved from http://www.jacksonville.com/news/20180309/what-makes-good-prosecutor-new-study-of-melissa-nelsons-office-hopes-to-find-out.
141. *See* https://measuresforjustice.org/.
142. *See* https://www.search.org/solutions/criminal-history-records/.
143. For an empirical comparison, *see, e.g.*, Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2018). A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint*. arXiv:1802.04422. *See also* Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3995–4004; Hajian, S., & Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE transactions on knowledge and data engineering, 25*(7):1445– 1459; Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268.
144. 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.
145. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & and Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd innovations in theoretical computer science conference, ACM*, 214–226; Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM*, 259–268; Price, E., Hardt, M., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in neural information processing systems. *arXiv*. Retrieved from https://arxiv.org/abs/1610.02413.
146. Varghese, R. (February 28, 2019). America's cities are running on software from the '80s. *Bloomberg Businessweek*. Retrieved from https://www.bloomberg.com/news/articles/2019-02-28/america-s-cities-are-running-on-software-from-the-80s?t=1551458437.
147. Wiseman, J. M. (2018). *Data-driven government: The role of chief data officers*. IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/data-driven-government-role-chief-data-officers.
148. Wiseman, J. M. (2018). *Data-driven government: The role of chief data officers*. IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/data-driven-government-role-chief-data-officers.
149. Ho, A. T.-K., & McCall, B. (2016). *Ten actions to implement big data initiatives: A study of 65 cities.* IBM Center for the Business of Government. Retrieved from http://www.businessofgovernment.org/report/ten-actions-implement-big-data-initiatives-study-65-cities.
150. Data scientists should be considered as part of the technical experts. We do not specifically mention "data scientists," as we view them as possessing skills that adept social scientists and computer scientists should possess.

151. Kutateladze, B., Lynn, V., & Liang, E. (2012). Do race and ethnicity matter in prosecution? A review of empirical studies. *Vera Institute of Justice*.
152. Luca, M., Kleinberg, J., & Mullainathan, S. (Jan-Feb 2016). Algorithms need managers, too. *Harvard Business Review*. Retrieved from https://hbr.org/2016/01/algorithms-need-managers-too.
153. Campolo, A., Sanfilippo, M., Whittaker, M., & Crawford, K. (2017). AI Now 2017 report. *AI Now Institute*. Retrieved from https://ainowinstitute.org/AI_Now_2017_Report.pdf.
154. *See* 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.
155. Desmarais, S., Garrett, B., & Rudin, C. (July 19, 2019). Risk assessment tools are not a failed 'Minority Report'. *Law360*. Retrieved from https://www.law360.com/.
156. *See* Corbett-Davies, S., Pierson, E., Feller, A., Soel, 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). New York: The Association for Computing Machinery; Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29* (pp. 3315-3323). Red Hook, NY: Curran Associates; Kamiran, F., Zliobaite, I., & Calders, T. (2013). Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems, 35*(3), 613-644.
157. Bolukbasi, T., Chang, K., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *CoRR*; Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268.
158. Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning - Volume 28*, pp. 325–333. *JMLR*.
159. Kamiran, F., & Calders, T. (2009). Classifying without discriminating. *IEEE Xplore Conference: Computer, Control, and Communication*. doi: 10.1109/IC4.2009.4909197
160. Johndrow, J., & Lum, K. (2019). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *Annals of Applied Statistics, 13*(1), 189-220.
161. 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.
162. Small/low data, or even the complete absence of data about certain subsets, often go unnoticed until algorithms are put into place. For this reason, data collection and analysis could help to identify such elisions and begin to correct for them. For example, such techniques have been used to identify neighborhoods in which individuals, for a variety reasons, underreport or fail to report shootings. *See* Griffiths, S. (April 19, 2016). Fighting a losing battle? AI ShotSpotter computer used to track gunfire reveals far more shots are fired than are ever reported. *Daily Mail*. Retrieved from https://www.dailymail.co.uk/sciencetech/article-3547719/Fighting-losing-battle-AI-ShotSpotter-computer-used-track-gunfire-reveals-far-shots-fired-reported.html.
163. Joseph, M., Kearns, M. J., Morgenstern, J., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29*, pp. 325–333.
164. Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy, 3*(3), 206–231.
165. Mancuhan, K., & Clifton, C. (2014). Combating discrimination using Bayesian networks. *Artificial Intelligence and Law, 22*(2), 211-238.
166. Lipton, Z. C., Chouldechova, A., & McAuley, J. (2017). Does mitigating ML's disparate impact require disparate treatment? *arXiv preprint*. arXiv:1711.07076.
167. Adeli, E., Zhao, Q., Pfefferbaum, A., Sullivan, E. V., Fei-Fei, L., Niebles, J. C., & Pohl, K. M. (2019). Bias-resilient neural network. *arXiv*, 1910.03676. Retrieved from https://arxiv.org/pdf/1910.03676.pdf.
168. 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; 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

169. Angwin, J. & Larson, J. (2016). Bias in criminal risk scores is mathematically inevitable, researchers say. *ProPublica*.
170. Bechavod, Y., & Ligett, K. (2018). Penalizing unfairness in binary classification. Retrieved from https://arxiv.org/pdf/1707.00044.pdf.
171. *See* https://18f.gsa.gov/.
172. Harcourt, B. E. (2017). *Against prediction: Profiling, policing, and punishment in an actuarial age*. Chicago: University of Chicago Press.
173. Ho, D. E. (2018). Judging statistical criticism. *Observational Studies, 4*, 42–56.
174. West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems: Gender, race, and power in AI. *AI Now Institute*. Retrieved from https://ainowinstitute.org/discriminatingsystems.html.
175. The University of Washington's Tech Policy Lab has developed a methodology for encouraging such deliberation. *See* https://techpolicylab.uw.edu/project/diverse-voices/.
176. *See* 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.
177. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv:1808.00023
178. Partnership on AI. (2019). *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/
179. Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (Oct. 2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*. Retrieved from https://hbr.org/2016/10/noise.
180. *See* Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (Apr. 2017). Creating simple rules for complex decisions. Harvard Business Review. Retrieved from https://hbr.org/2017/04/creating-simple-rules-for-complex-decisions.
181. Jung, J., Concannon, C., Shroff, R., Goel, S., & Goldstein, D. G. (2017). Simple rules for complex decisions. *arXiv*. Retrieved from https://arxiv.org/abs/1702.04690.
182. Kahneman, D., Rosenfield, A. M., Gandhi, L., & Blaser, T. (Oct. 2016). Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard Business Review*. Retrieved from https://hbr.org/2016/10/noise.
183. Heaton, B. (May 15, 2015). New York City fights fire with data. *Government Technology.* Retrieved from https://www.govtech.com/public-safety/New-York-City-Fights-Fire-with-Data.html.
184. Garrett, B. L., & Monahan, J. (2018). Judging risk. *Virginia Public Law and Legal Theory Research Paper No. 2018-44*.
185. Main, F. (July 3, 2016). Cook County judges not following bail recommendations: study. *Chicago Sun-Times*. Retrieved from https://chicago.suntimes.com/2016/7/3/18325456/cook-county-judges-not-following-bail-recommendations-study.
186. *See* Garrett, B. L., & Monahan, J. (2018). Judging risk. *Virginia Public Law and Legal Theory Research Paper No. 2018-44*; Yang, C. S. (2017). Toward an optimal bail system. *New York University Law Review, 92*, 1399-1493.
187. See Garrett, B. L., & Monahan, J. (2018). Judging risk. *Virginia Public Law and Legal Theory Research Paper No. 2018-44*; Hilton, N. Z., Scurich, N., & Helmus, L.-M. (2015). Communicating the risk of violent and offending behavior: Review and introduction to this special issue. *Behavioral Sciences & the Law, 33*, 1-18.
188. Fountaine, T., McCarthy, B., & Saleh, T. (July-August, 2019). Building the AI-powered organization. *Harvard Business Review*, 62-73.

189. Dawes, R. M., Faust, D. & Meehl, P. E. (1989). Clinical vs. actuarial judgment. *Science, 243*, 1668-1674; Gendreau, P., Little, T. & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34*, 575-607; 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.
190. 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/publications/beyond-algorithm.
191. *See* Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review, 66*, 803-872; Oleson, J. C. (2011). Risk in sentencing: Constitutionally suspect variables and evidence-based sentencing. *Southern Methodist University Law Review, 64*, 1329-1402.
192. Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal, 68*(6), 1043-1134. However, *Ricci v. DeStefano*, 557 U.S. 55, 585 (2009) allows for such use (in this case, the Supreme Court held that an employer could engage in intentional discrimination if it did so to avoid being subject to liability arising from current racially disparate impact).
193. See Desmarais, S., Garrett, B., & Rudin, C. (July 19, 2019). Risk assessment tools are not a failed 'Minority Report'. *Law360*. Retrieved from https://www.law360.com; 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/publications/beyond-algorithm.
194. *See* Coglianese, C., & Lehr, D. (2019). Transparency and algorithmic governance. *Administrative Law Review, 71*, 1-56; Coglianese, C., & Lehr, D. (2017). Regulating by robot. *Georgetown Law Journal, 105*, 1147-1223.
195. Henke, N., Levine, J., & McInerney, P. (2018). You don't have to be a data scientist to fill the must-have analytics role. *Harvard Business Review*. Retrieved from https://hbr.org/2018/02/you-dont-have-to-be-a-data-scientist-to-fill-this-must-have-analytics-role?autocomplete=true.
196. Fountaine, T., McCarthy, B., & Saleh, T. (July-August, 2019). Building the AI-powered organization. *Harvard Business Review*, 62-73.
197. Fountaine, T., McCarthy, B., & Saleh, T. (July-August, 2019). Building the AI-powered organization. *Harvard Business Review*, 62-73.
198. Kutateldze, B. (forthcoming). 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.
199. Clarke, K. A. (2005). The phantom menace: Omitted variable bias in econometric research. *Conflict Management and Peace Science, 22*(4):341–352.
200. 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.
201. Huq, A. Z. (2019). Racial equity in algorithmic criminal justice. *Duke Law Journal, 68*(6), 1043-1134.
202. 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.
203. Bechavod, Y., & Ligett, K. (2018). Penalizing unfairness in binary classification. Retrieved from https://arxiv.org/pdf/1707.00044.pdf.
204. Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery, 21*(2):277–292; Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268.
205. 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 that cannot be meaningfully transformed, and prosecutors would have to analyze the cases without machine guidance.

ABOUT THE AUTHORS

Joseph J. Avery is a Graduate Fellow of the Woodrow Wilson Scholars at Princeton University. His research is funded by the National Defense Science and Engineering Graduate Fellowship (Department of Defense). He holds a B.A. in philosophy and economics from NYU, a M.A. in psychology from Princeton, and a J.D. from Columbia Law School. His teaching and research interests center around criminal law, criminal procedure, and evidence, with related interests in legal technology, including the development of novel computational law tools.

JOSEPH J. AVERY

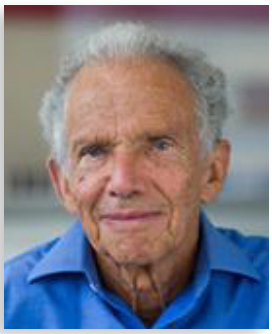
Akbar A. Agha is a Researcher at the Princeton Project in Computational Law. He concurrently works in the Maintenance, Repairs, and Operations industry as a Data Scientist for W.W. Grainger Inc. He received his B.Sc. in Mechanical Engineering from the University of Illinois at Urbana-Champaign. His research interests include bias in machine learning models, regulations and incentives in the development of machine learning tools, and AI policy.

AKBAR A. AGHA

Eric J. Glynn is a Doctoral Candidate at Princeton University working in the Departments of Molecular Biology and Computer Science. His research is funded by the National Institute of Health Graduate Research Fellowship (NIH). He holds a B.Sc. in Biology and Chemistry from Alma College (Alma, MI) and a M.A. in Molecular Biology from Princeton. His research focuses on applications of Machine Learning to open questions in human health and Molecular Biology; specifically, building tools to understand and predict the body's immune response to cancer.

ERIC J. GLYNN

Joel Cooper is Professor of Psychology at Princeton University, having joined the faculty after receiving his Ph.D. from Duke University. He has served as chair of the Princeton Psychology Department and is former editor of the Journal of Experimental Social Psychology. He is the author of several books and a co-editor of the Sage Handbook of Social Psychology. His research interests include the study of social attitudes, cognitive dissonance and the impact of expert witness testimony.



JOEL COOPER

KEY CONTACT INFORMATION

To contact the authors:

Joseph J. Avery

522 Peretsman Scully Hall
Department of Psychology
Princeton University
Princeton, NJ 08540

Phone: (609) 258-7344

javery@princeton.edu

Akbar A. Agha

522 Peretsman Scully Hall
Department of Psychology
Princeton University
Princeton, NJ 08540

akbar.agha0712@gmail.com

Eric J. Glynn

522 Peretsman Scully Hall
Department of Psychology
Princeton University
Princeton, NJ 08540

ejglynn@princeton.edu

Joel Cooper

522 Peretsman Scully Hall
Department of Psychology
Princeton University
Princeton, NJ 08540

jcoops@princeton.edu

The Princeton Project in Computational Law
Princeton, NJ