HARVARD UNIVERSITY

Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Harvard John A. Paulson School of Engineering and Applied Sciences
have examined a dissertation entitled:

"Risky Reforms: A Sociotechnical Analysis of Algorithms as Tools for Social Change"

presented by: Ben Green

candidate for the degree of Doctor of Philosophy and here by

Signature _____
Typed name: Professor Y. Chen

Signature _____
Typed name: Professor F. Doshi-Velez

Signature _____
Typed name: Professor M. Gray

Signature: _____
Typed name: Professor A. Papachristos

August 25, 2020

Risky Reforms: A Sociotechnical Analysis of Algorithms as Tools for Social Change

a dissertation presented
by
Ben Green
to
The School of Engineering and Applied Sciences

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Applied Mathematics

Harvard University
Cambridge, Massachusetts
August 2020

Risky Reforms: A Sociotechnical Analysis of Algorithms as Tools for Social Change

Abstract

This thesis considers the relationship between efforts to address social problems using algorithms and the social impacts of these interventions. Despite widespread optimism about algorithms as tools to promote reform and improve society, there is often little rigorous analysis regarding how algorithmic interventions will lead to particular desired outcomes. In turn, many well-intentioned applications of algorithms have led to social harm. In this thesis, I focus on the use of "risk assessments" in the U.S. criminal justice as a notable example of machine learning algorithms being used as tools for social change. Treating these algorithmic interventions as sociotechnical and political reform efforts rather than primarily technical projects, I center my analyses of risk assessments around their social and political consequences. In Part I (Interaction), I introduce a new "algorithm-in-the-loop" framework for evaluating the impacts of algorithms in practice, using experiments to uncover unexpected behaviors that occur when people collaborate with risk assessments. In Part II (Risk and Response), I interrogate typical conceptions of risk and how to respond to it, developing a novel machine learning method to analyze structural factors of violence and to support non-punitive and public health-inspired violence prevention efforts. In Part III (Reform), I place these technical studies in the broader context of social and political reform, describing the limits of risk assessments as a tool for criminal justice reform and articulating a new mode of practice—"algorithmic realism"—that synthesizes computer science, law, STS, and political theory in order to equip computer scientists to work more rigorously in the service of social change. By expanding the scope of questions asked of risk assessments, this dissertation sheds new light on how risk assessments represent a "risky" strategy for achieving criminal justice reform. Through these analyses, I chart the beginnings of a more interdisciplinary and rigorous approach to evaluating and developing algorithms as tools for social change.

Contents

Acknowledgments

I first want to thank my advisor, Yiling Chen, for all of her support and mentorship. Yiling graciously took me on as a student halfway through my PhD and has been an incredible and empowering advisor, teaching me how to develop a research agenda, supporting all of my interdisciplinary interests and side projects, funding me to attend conferences and workshops, and helping me navigate the job market. I have learned a great deal from her about how to be a good scholar, collaborator, and advisor, and look forward to calling on her wisdom and advice for many years.

I am also grateful to the rest of my committee. Andrew Papachristos has been a wonderful mentor and collaborator dating back to my undergraduate days. I have learned so much from Mary Gray's commitment to rigorous interdisciplinary scholarship and the deep care for others that she brings to her work. Finale Doshi-Velez inspires me with her ability to bridge machine learning with other disciplines and her steadfast support for her students.

I also want to thank my initial advisor, Radhika Nagpal. Radhika taught me a great deal about how to conduct research and lead a research group during my first few years in graduate school. She responded with remarkable grace and support as my research interests diverged from hers and has continued to be a trusted mentor.

I am grateful to the many friends who have made my years at Harvard so fun and intellectually stimulating, in particular Rediet Abebe, Sarah Balakrishnan, Eric Balkanski, Blake Dickson, Will Holub-Moorman, Thibaut Horel, Lily Hu, Jean Pouget-Abadie, Berk Ustun, and Zach Wehrwein. The Berkman Klein Center for Internet & Society and the Harvard Graduate Students Union have also been wonderful hubs of friendship and inspiration. My time at Harvard would have been far less fun and interesting without you.

My parents, brothers, and grandparents have supported me and my academic pursuits for as long as I can remember, from providing math lessons and book recommendations when I was young to reviewing paper drafts today. I am grateful to have been able to spend my graduate school years so close to my family.

Salomé Viljoen has been a wellspring of ideas, inspiration, and support. This dissertation would be nowhere near as fun to have produced nor interesting to read if not for her.

The work in this thesis is based on joint work with Yiling Chen, Thibaut Horel, Andrew Papachristos, and Salomé Viljoen and was supported by the National Science Foundation. I have also been fortunate to learn from and collaborate with Paul Bardunias, Lily Hu, J. Scott Turner, Radhika Nagpal, and Justin Werfel.

Chapter 1

Introduction

Following recent advances in the quality and accessibility of machine learning algorithms, the public sector increasingly uses machine learning as a central tool to distribute resources and make important decisions [194]. Similarly, much of the work in computer science labs and technology companies is motivated by a desire to improve society. Many computer scientists aim to "change the world" [378], leading to the development of algorithms for use in courts [17], city governments [194], hospitals [468], schools [526], and other essential societal institutions. A particularly common goal among algorithm developers is to contribute to the "social good," with countless such efforts among academic institutes, conferences, companies, and volunteer organizations [191].

Despite this optimism about the value of algorithms as tools to promote reform and improve society, there is often little articulation or rigorous analysis of how algorithmic interventions will lead to particular desired outcomes. Even as algorithms are hailed for their ability to improve society, they are predominantly evaluated along traditional technical metrics such accuracy and efficiency. Thus, alongside computer science's growing interest in addressing social challenges has come a recognition—driven by affected communities and scholarship in science, technology, and society (STS) and critical algorithm studies—that many well-intentioned applications

of algorithms have led to harm. Algorithms can be biased [17], discriminatory [27], dehumanizing [365], and violent [229, 373]. They can exclude people from receiving social services [154, 369], spread hateful ideas [421, 499], and facilitate government oppression of minorities [305, 353].

This gap between the intentions behind the development and use of algorithms and the social impacts of many algorithms raises central questions at the heart of efforts to improve society: What is the relationship between social interventions and the impacts of those interventions? How can social interventions robustly generate their desired social impacts? Scholars across a wide range of fields have long noted that political reforms often fail to achieve their desired goals and can have unintended adverse consequences [5, 109, 316, 354, 483]. Many have also noted how technological innovations and applications can fail to achieve their desired goals and lead to unexpected social harms [13, 194, 252, 441, 479, 512].

This thesis looks to the relationship between algorithmic interventions (efforts to use algorithms to address social problems) and social impacts. I focus on the use of "risk assessments" in the U.S. criminal justice system (particularly in courts, with some discussion of policing in Part II) as a notable example of machine learning algorithms being used as a tool for social change. Treating these algorithmic interventions as sociotechnical and political reform efforts rather than primarily technical projects, I center my development and analysis of risk assessment algorithms around their social and political consequences. By "sociotechnical," I refer to the ways in which social actors and technological artifacts become intertwined as part of unified—rather than discrete—systems (or networks) and the mode of analysis that analyzes technology in relation to these social actors [290, 469]. Within sociotechnical systems, "technologies can be assessed only in their relations to the sites of their production and use" [469].

In Part I (Interaction), I introduce a new "algorithm-in-the-loop" framework for evaluating the impacts of algorithms in practice, using experiments to uncover unexpected behaviors that occur when people collaborate with risk assessments. In Part II (Risk and Response), I interrogate typical conceptions of risk and how to respond to it, developing a novel machine learning method to analyze structural factors of violence and to support

non-punitive and public health-inspired violence prevention efforts. In Part III (Reform), I place these technical studies in the broader context of social and political reform, describing the limits of risk assessments as a tool for criminal justice reform and articulating a new mode of practice—"algorithmic realism"—that synthesizes computer science, law, STS, and political theory in order to equip computer scientists to work more rigorously in the service of social change. By expanding the scope of questions asked of risk assessments, this dissertation sheds new light on how risk assessments represent a "risky" strategy for achieving criminal justice reform. Through this process, however, I chart the beginnings of a more interdisciplinary and rigorous approach to evaluating and developing algorithms as tools for social change.

1.1 Criminal Justice Risk Assessments

Across the United States, many oft-opposed groups have united around risk assessments as a promising path forward for adjudication in criminal courts: Democrats and Republicans [217], conservative states [255] and liberal states [361], criminal defense organizations [406] and prosecutors [339]. In turn, risk assessments have proliferated in recent years: in 2017, 25% of people in the U.S. lived in a jurisdiction using a pretrial risk assessment, compared to just 10% four years prior [403]. A 2019 scan of 91 U.S. jurisdictions found that more than two-thirds used a pretrial risk assessment [404].

Risk assessments are mechanisms for identifying potential risks, the likelihood of those risks manifesting, and the consequences of those events [413]. Within the criminal justice system, risk assessments are most widely used in the contexts of pretrial detention (to predict the likelihood that a criminal defendant will fail to appear in court for trial and, in some jurisdictions, will commit a crime before trial) and sentencing and parole (to predict the likelihood that a defendant or inmate will commit a crime in the future). Risk assessments are also used in policing to predict places where crime is likely to occur and people likely to be the perpetrators or victims of violence [241]. Although actuarial risk assessments have existed within the criminal justice system for several

decades, today's tools represent a new generation that incorporates a larger range of risk factors and is often developed through more advanced statistical methods (such as machine learning) [46, 266].

The recent push toward adopting risk assessments is largely motivated by the criminal justice system's current crisis of legitimacy. Scholarship and activism have demonstrated the countless ways in which racism is baked into the criminal justice system's fundamental structure [10, 105, 228, 354, 462]. Through popular books about mass incarceration [10], racial justice movements such as Black Lives Matter, and increased attention to the inequity of policies such as cash bail [106], there is a growing consensus that the criminal justice system is rife with discrimination. Even criminal justice system actors and defenders have acknowledged the need for change. In 2015, more than 130 police chiefs and prosecutors formed a new organization to combat mass incarceration [169]; the following year, the largest police organization in the U.S. apologized for policing's "historical mistreatment of communities of color" [246]. More recently, politicians (including former prosecutors) who formerly embraced "tough on crime" policies have apologized for their actions and championed criminal justice reform [156, 183, 310].

Risk assessments are often hailed as an important tool for addressing some of the central issues in pretrial and sentencing adjudication. The theory of change regarding the benefits of risk assessments is grounded in two key assumptions. The first assumption is that risk assessments will mitigate judicial and policing biases by providing "objective" decisions [115, 217, 257, 361, 366, 495].

The second assumption is that risk assessments will promote criminal justice reform. This is expected to occur through objective risk assessments replacing discriminatory policies and reducing incarceration. For example, Senators Kamala Harris and Rand Paul introduced the Pretrial Integrity and Safety Act of 2017, proposing to replace money bail with risk assessments so that pretrial release would be based on risk rather than wealth and so that pretrial release rates would increase [217]. Several states have implemented pretrial risk assessments with these same goals [255, 361]. Many endorsements of evidence-based sentencing are similarly grounded in the goal of reducing incarceration [346, 460]. Predictive policing systems are similarly adopted as a tool for policing

reform [161].

Supporters of risk assessments draw a clear link between objectivity and reform. In its Statement of Principles, Arnold Ventures (the organization behind the Public Safety Assessment (PSA), a pretrial risk assessment used in nineteen states [293]) writes that the goal of its criminal justice reform efforts is to promote "a criminal justice system that dramatically reduces the use of pretrial detention." Developing the PSA was one of its "earliest investments in pretrial reform," under the belief that "[p]roviding judges with an objective means to consider only relevant data may counterbalance some [human] biases and lead to fairer pretrial outcomes" [495]. Similarly, the Attorney General of New Jersey described the state's adoption of "an objective pretrial risk-assessment" as "[o]ne of the most critical innovations undergirding the entire [statewide bail] reform initiative" [393].

1.2 Algorithmic Fairness

Given that a primary motivation behind risk assessments is to promote objective and unbiased decision-making, significant debate has focused on how to measure and ensure that risk assessments do not discriminate against Blacks. The satisfaction of statistical metrics for fairness has become a central component of evaluating the objectivity of risk assessments.

In May 2016, the investigate journalism outlet ProPublica released a report exposing the racial bias in computer algorithms known that "risk assessments" that predict future criminality [17]. ProPublica obtained the risk scores of more than 7,000 pretrial defendants in Broward County, Florida in 2013 and 2014, predictions which had been generated by an algorithm known as COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) created by the company Northpointe (since renamed as Equivant). The team's analysis of these predictions found that the algorithm is "biased against blacks." Blacks were disproportionately mislabeled as "high risk" (i.e., were subject to false positive predictions) and whites were disproportionately mislabeled as "low risk" (i.e., were subject to false negative predictions). For instance, 45% Blacks who did not recidivate over the

two-year evaluation period were labeled "high risk," compared to just 23% of whites who did not recidivate [289].

ProPublica was not the first to suggest that algorithms could be biased. Scholarship and policy across a range of areas had long considered how algorithms, software, and quantitative assessments could discriminate [172, 236, 239, 274, 370]. Such concerns gained particular traction as big data and algorithms become more pervasive in the criminal justice system and many aspects of daily life [27, 28, 144, 364, 455, 473]. In 2014, Attorney General Eric Holder expressed concern that criminal justice risk assessments "may exacerbate unwarranted and unjust disparities" [232].

Although some responded to ProPublica's article with outrage about racist algorithms making criminal justice decisions [134, 377, 456], many critiqued ProPublica's conclusion. Northpointe responded that ProPublica "did not present any valid evidence that the risk scales are biased against blacks" [130]. In particular, Northpointe argued that ProPublica failed to account for the different recidivism rates between Blacks and whites and mistakenly focused on false positive and false negative rates as evidence of racial bias. Using the "predictive parity" measure that they assert is the proper evaluation metric, Northpointe demonstrated that COMPAS's predictions of high and low risk were both equally accurate across race and therefore unbiased. Northpointe's response was not simply a corporation shamelessly defending itself against public scrutiny. Many independent researchers similarly rejected claims that COMPAS is racially biased on the grounds that, at each level of predicted risk, Blacks and whites were approximately equally likely to recidivate [100, 166, 185].

Several groups of researchers soon showed that the conflict between these different notions of bias within the COMPAS algorithm were mathematically inevitable [16, 82, 278]. These results, known as the "impossibility of fairness," concern two categories of statistical metrics for fairness.

The first category involves metrics emphasizing equal treatment. Although there are subtle differences between these metrics, they all express a similar logic that similar people should be treated similarly, with similarity based on the algorithm's predictions. These metrics include:

- Calibration: predictions of risk should reflect the same underlying level of risk across groups [82, 278].
- Predictive parity: outcome rates among people labeled "high risk" should be the same across groups [130].
- Threshold rules: decisions should be based on a single risk threshold for all groups [98].

The second category involves metrics emphasizing the impacts of prediction errors. These include:

- Error rate balance: false positive and false negative rates should be equal across groups [82].
- Balance for the positive/negative class: the average score assigned to positive/negative predictions should be the same across groups [278].

Given these two sets of metrics, the impossibility of fairness shows that if two groups have different rates of an outcome, then it is impossible for predictions about those groups to both be calibrated and have balanced errors [82, 278]. Both types of fairness measures can be simultaneously achieved only in two unlikely scenarios: when the algorithm is able to perfectly predict each person's outcome or when the two groups have identical outcome rates [278]. In the context of risk assessments, this means that, given higher crime rates among Black defendants than white defendants, it is impossible for a risk assessment to make calibrated predictions of risk without having a higher false positive rate and lower false negative rate for Black defendants.

1.3 Outline

Here I provide a brief outline for the dissertation, with references to the individual works from which each chapter is adapted.

Part I (Interaction) presents three studies regarding the role of human-algorithm interactions in the implementation of algorithms. Although significant debate has centered on the statistical properties of risk assessments, little research has considered how these algorithms affect the people who actually make decisions. In practice,

risk assessments do not make decisions—they inform judges, who are the final arbiters. Chapter 2 introduces a new framework—"algorithm-in-the-loop" decision making—that emphasizes the human's decisions as the central output of importance when risk assessments are used [196, 197, 198]. Chapter 3 uncovers racial biases in human responses to risk assessments (what I call "disparate interactions") [196]. Chapter 4 studies the failure of algorithm-in-the-loop decision making to satisfy core tenets of just decision making [197]. Chapter 5 demonstrates how using algorithmic risk assessments can increase risk aversion as a decision-making factor in government contexts [198].

Part II (Risk and Response) reframes risk prediction around social structures and social services. Chapter 6 describes the rise of "predictive policing" algorithms and the need to center policy reforms rather than simply make existing practices more efficient [194]. Chapter 7 introduces a novel machine learning method to study the spread of gunshot victimization through social networks and to support non-punitive and public health-inspired violence prevention efforts [199].

Part III (Reform) considers the implications of these technical studies for projects of social reform. Chapter 8 demonstrates the limits of risk assessments' supposed objectivity and argues instead that risk assessments are likely to legitimize rather than combat the criminal justice system's carceral logics and policies [195]. Chapter 9 proposes a new mode of practice—"algorithmic realism" that can equip computer scientists to engage more rigorously with the sociality of their work and to develop interventions that account for rather than reproduce historical injustice [201]. Chapter 10 briefly concludes by outlining some salient areas of future work related to algorithms and social change.

Part I

Interaction

Chapter 2

An Algorithm-in-the-Loop Approach to Decision-Making

2.1 Introduction

People and institutions increasingly make important decisions with the aid of risk assessments. Applications of algorithmic risk assessments include directing police and social services to individuals most at risk of being involved in gun violence [434], informing pretrial and sentencing decisions with a criminal defendant's likelihood to recidivate [17, 361], targeting public health inspections based on the risk of illness [194, 397], and predicting which children are most likely to be abused or neglected [154]. Applications of risk assessments in other contexts include banks using models to manage credit risk [450]. All of these settings involve machine learning models that inform people who are tasked with making decisions. This trend represents a fundamental shift in decision-making: where in the past decision-making was a social enterprise, decision-making today has become a sociotechnical affair.

This shift in decision-making is particularly notable with respect to high-stakes settings such as courts within the U.S. criminal justice system. Across the United States, courts are increasingly using risk assessments to estimate the likelihood that criminal defendants will engage in unlawful behavior in the future. These tools are being deployed during several stages of criminal justice adjudication, including at bail hearings (to predict the risk that the defendant, if released, will be rearrested before trial or not appear for trial) and at sentencing (to predict the risk that the defendant will recidivate). Because risk assessments rely on data and a standardized process, many proponents believe that they can mitigate judicial biases and make "objective" decisions about defendants [217, 361, 115]. Risk assessments have therefore gained widespread support as a tool to reduce incarceration rates and spur criminal justice reform [406, 217, 361].

Yet many are concerned that risk assessments make biased decisions due to the historical discrimination embedded in training data. For example, the widely-used COMPAS risk assessment tool wrongly labels Black defendants as future criminals at twice the rate it does for white defendants [17]. Prompted by these concerns, machine learning researchers have developed a rapidly-growing body of technical work focused on topics such as characterizing the incompatibility of different fairness metrics [278, 82] and developing new algorithms to reduce bias [158, 215].

Despite these efforts, current research into fair machine learning fails to capture an essential aspect of how risk assessments impact decision-making in courts: their influence on judges. After all, risk assessments do not make definitive decisions about pretrial release and sentencing—they merely aid judges, who must decide whom to release before trial and how to sentence defendants after trial. In other words, algorithmic outputs act as decision-making aids rather than final arbiters. Thus, whether a risk assessment *itself* is accurate and fair is of only indirect concern—the primary considerations are how it affects decision-making processes and whether it makes *judges* more accurate and fair. No matter how well we characterize the technical specifications of risk assessments, we will not fully understand their impacts unless we also study how judges interpret and use them.

The chain from algorithm to person to decision has become vitally important as algorithms inform increasing

numbers of high-stakes decisions. To improve our understanding of these contexts, I introduce an "algorithm-in-the-loop" framework that places algorithms in a sociotechnical context—thus focusing attention on human-algorithm interactions to improve human decisions rather than focusing on the algorithm to improve its decisions. Rigorous studies of algorithm-in-the-loop systems are necessary to inform the design and implementation of algorithmic decision-making aids being deployed in courts and beyond.

After situating the algorithm-in-the-loop framework within the literature, I describe my experimental approach to evaluating algorithm-in-the-loop systems. The following three chapters describe a series of studies using this approach. The results of these studies highlight the urgent need to more rigorously study the impacts of risk assessments, focusing on the full set of mechanisms through which potential outcomes come to pass. Risk assessments have the potential to improve decision-making, but can also lead to unintended outcomes as they are integrated into human decision-making processes and broader political contexts; evaluations must therefore be grounded in rigorous sociotechnical analyses of the downstream impacts. As the following studies indicate, one essential component that shapes these outcomes is the quality and reliability of human-algorithm interactions. Continued research into how people should and do collaborate with machine learning models is necessary to inform the design, implementation, and governance of algorithmic decision-making aids being deployed across society.

2.2 Related Work

A core component of integrating a technical system into social contexts is ensuring that people recognize when to rely on the tool and when to discount it. As technology is embedded into critical human decisions, the stakes of human trust and reliance on technology rise, such that "poor partnerships between people and automation will become increasingly costly and catastrophic" [295]. Recent breakdowns in the human-automation partnership in self-driving cars and airplane autopilot have led to disaster [221, 40]. In many contexts, designing effective

human-machine collaborations hinges as much (if not more) on presenting guidance that is tailored to human trust and understanding as it does on providing the technically optimal advice [295, 150].

Significant research in human-computer interaction has considered how to develop systems that effectively integrate human and computer intelligence [262, 237]. In the context of algorithm-assisted human decision-making, prior research has explored topics such as what interactions can facilitate the development of machine learning models [285, 155, 74], how to improve human performance with an algorithm's assistance [89, 287], and the ways in which laypeople perceive algorithmic decisions [296, 39, 152]. Research related to human-algorithm interactions when making predictions can be summarized into two broad categories of findings.

2.2.1 People struggle to interpret and effectively use algorithms when making decisions

The phenomenon of "automation bias" suggests that automated tools influence human decisions in significant, and often detrimental, ways. Two types of errors are particularly common: omission errors, in which people do not recognize when automated systems err, and commission errors, in which people follow automated systems without considering contradictory information [350]. Heavy reliance on automated systems can alter people's relationship to a task by creating a "moral buffer" between their decisions and the impacts of those decisions [112]. Thus, although "[a]utomated decision support tools are designed to improve decision effectiveness and reduce human error, […] they can cause operators to relinquish a sense of responsibility and subsequently accountability because of a perception that the automation is in charge" [112].

Several experimental studies have uncovered important issues that arise when people use algorithms to inform their decision-making. People often discount algorithmic recommendations, preferring to rely on their own or other people's judgment and exhibiting less tolerance for errors made by algorithms than errors made by other people [520, 302, 131]. This may be due in part to the fact that people struggle to evaluate their own and the algorithm's performance [287]. Although people appear in some contexts to follow correct predictions

more than incorrect ones [287], other studies suggest that people are unable to distinguish between reliable and unreliable predictions [186], to deviate correctly from algorithmic forecasts [132], or to detect algorithmic errors [400]. Moreover, people have been shown to be influenced by irrelevant information, to rely on algorithms that are described as having low accuracy, and to trust algorithms that are described as accurate but actually present random information [287, 459, 151]. People with more expertise are less willing than laypeople to take advice from actuarial or algorithmic sources [309, 247, 334] And despite widespread calls for explanations and interpretable models, recent studies have found that simple models do not lead to better human performance than Black box models [400] and that varying algorithmic explanations does not affect human accuracy [357].

2.2.2 People often use algorithms in unexpected and biased ways

Previous research suggests that information presumed to help people make fairer decisions can fail to do so because it filters through people's preexisting biases. For example, "ban-the-box" policies (which are intended to promote racial equity in hiring by preventing employers from asking job applicants whether they have a criminal record) actually increase racial discrimination by allowing employers to rely on stereotypes and thereby overestimate how many Black applicants have criminal records [5, 135]. Similarly, people's interpretations of police-worn body camera footage are significantly influenced by their prior attitudes about police [458].

A particular danger of breakdowns in human-algorithm collaborations is that the application of an algorithm will lead to unintended behaviors and decisions. Ethnographic studies have documented how the uses of algorithms in practice can differ significantly from the planned and proclaimed uses, with algorithms often being ignored or resisted by those charged with using them [84, 54]. In other cases, the application of algorithms has prompted people to alter their behavior, becoming overly fixated on the algorithm's advice or focusing on different goals [45, 238].

Pretrial risk assessments represent a notable example of algorithms that are highly indeterminate and often do not generate the intended or expected results. Studies have shown that judges harbor implicit biases and that racial

disparities in incarceration rates are due in part to differential judicial decisions across race [409, 2]. In Florida, for example, white judges give harsher sentences to Black defendants than white ones who have committed the same crime and received the same score from the formula the state uses to set criminal punishments [428]. Although risk assessment algorithms are typically adopted with the explicit goal of reducing detention rates, in many cases they have had only negligible impacts because judges ignore the majority of recommendations for release. Risk assessments used in Kentucky and Virginia have thus far failed to produce significant and lasting increases in pretrial release, as judges often overrode the risk assessment when it recommended release and reduced their reliance on the risk assessment over time [464, 465]. Similar results have been found in Cook County, Illinois [319] and in Santa Cruz and Alameda County, California [503].

There is also evidence that people's interactions with risk assessments are fraught with racial biases. Similarly, analyses have observed that judges in Broward County, Florida penalized Black defendants more harshly than white defendants for crossing into higher risk categories [107] and that judicial use of a risk assessment in Kentucky increased racial disparities in pretrial outcomes [8].

2.3 The Algorithm-in-the-Loop Framework

As computational systems permeate everyday life and inform critical decisions, it is of paramount importance to study how algorithmic predictions impact human decision-making across a broad range of contexts. Risk assessments are just one of an emerging group of algorithms that are intended to inform people making decisions (other examples include predictions to help companies hire job applicants and to help doctors diagnose patients). Yet despite robust research into the technical properties of these algorithms, we have a limited understanding of their sociotechnical properties: most notably, whether and how they actually improve decision-making. To answer these questions, it is necessary to study algorithms following the notion of "technologies as social practice," which is grounded in the understanding that technologies "are constituted through and inseparable from

the specifically situated practices of their use" [469].

A natural body of work from which to draw inspiration in studying human-algorithm collaborations is human-in-the-loop (HITL) systems. In settings such as social computing and active learning, computational systems rely on human labor (such as labeling photos and correcting errors) to overcome limitations and improve their performance. But where HITL processes privilege models and algorithms, utilizing people where necessary to improve computational performance, settings like pretrial release operate in reverse, using algorithms to improve human decisions.

This distinction suggests the need for an alternative framework: algorithm-in-the-loop (AITL) systems (Figure 2.1).[1] Instead of improving computation by using humans to handle algorithmic blind spots (such as analyzing unstructured data), AITL systems improve human decisions by using computation to handle cognitive blind spots (such as finding patterns in large, complex datasets). This framework centers human-algorithm interactions as the locus of study and prioritizes the human's decision over the algorithm's as the most important outcome.

An algorithm-in-the-loop perspective can inform essential sociotechnical research into algorithms. Recent work related to interpretability provides one important direction where progress is already being made [400, 357, 136]. Future analysis should focus on how to develop and present algorithms so that people can most effectively and fairly incorporate them into their deliberative processes, with particular attention to improving evaluations of algorithm quality and reducing disparate interactions. This may involve altering the algorithm in unintuitive ways: previous research suggests that in certain situations a seemingly suboptimal algorithm actually leads to better outcomes when provided to people as advice [150].

It will also be important to study the efficacy of different mechanisms for combining human and algorithmic judgment across a variety of contexts. Most algorithm-in-the-loop settings involve simply presenting an algorithmic output to a human decision-maker, relying on the person to interpret and incorporate that information.

[1] Although previous studies have used the phrase "algorithm-in-the-loop," they have defined it in the context of simulation and modeling rather than in relation to human-in-the-loop computations and human-algorithm interactions [517, 433].

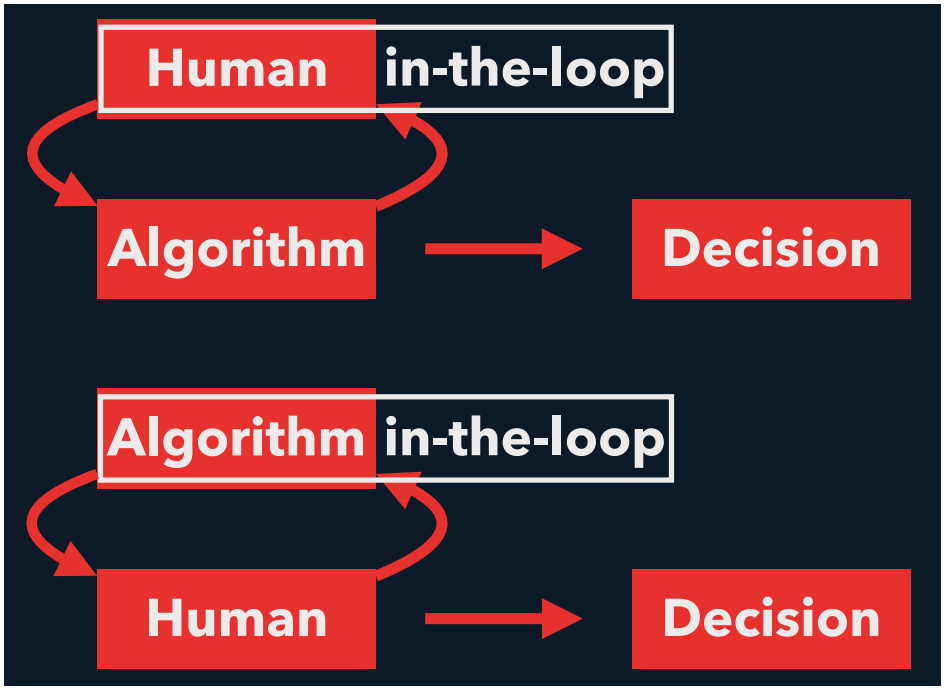

Figure 2.1: A visual representation of the distinction between HITL and AITL settings. While HITL settings center algorithms with humans providing aid, AITL settings center humans with algorithms providing aid.

Yet research within human-computer interaction and crowdsourcing suggests that alternative approaches could lead to a better synthesis of human and computer intelligence [89, 237, 261, 262]. Which mechanisms are most effective (and desirable from an ethical and procedural standpoint) will likely vary depending on the situation.

Finally, given that automation can induce a moral buffer [112], it is necessary to study how using algorithms affects people's sense of responsibility for their decisions. Given the all-too-common expressions from engineers that they do not bear responsibility for the social impacts of their technologies [191], the potential for automation bias raises the unsettling specter of situations in which both the engineers developing algorithms and the people using them believe the other to be primarily responsible for the social outcomes. It is of vital importance to study whether algorithms create a moral buffer and to find ways to avoid such scenarios.

The algorithm-in-the-loop framework and the studies that follow demonstrate the importance of an experimental and diagnostic approach to studying the impacts of risk assessments on government decision-making. Given evidence of algorithms producing unexpected impacts in practice [55, 466, 464], there is an urgent need to

uncover implementation issues *before* these algorithms are used to shape life-changing decisions. Experimental studies with laypeople present one promising approach for gaining this preliminary diagnostic knowledge about how algorithms are likely to affect human decisions in practice. Although the gold standard is empirical data on how expert decision-makers are influenced by risk assessments in practice, a controlled experimental setup enables insights that would be difficult to obtain in real-world settings and numerous experimental studies have observed behaviors among judges that resemble those of laypeople [207, 410]. Knowledge gained through such experiments can inform the development, implementation, and governance of algorithms in real-world settings and prevent the implementation of technical systems whose social impacts are untested.

2.4 Experimental Approach

Our study progressed in two stages. The first stage involved developing risk assessments for pretrial detention and financial lending. The second stage consisted of running experiments on Amazon Mechanical Turk to evaluate how people interact with these risk assessments when making predictions and decisions. The full study was reviewed and approved by the Harvard University Institutional Review Board and the National Archive of Criminal Justice Data.

2.4.1 Description of Settings

The studies described in the following chapters focused on two settings: pretrial release in the criminal justice system and financial lending from banks and the government. All three studies included the pretrial setting. The second study included a version of the lending setting focused on generic loans provided by a bank and the third study included a version of the lending setting focused on government home improvement loans.

Pretrial Release

When someone is arrested, courts can either hold that person (a "criminal defendant") in jail until their trial or release them with a mandate to return for their trial (many people are also released under conditions such as paying a cash bond or being subject to electronic monitoring). Among other considerations, courts aim to ensure that defendants will return to court for trial and will not commit any crimes if released. Jurisdictions across the United States have therefore turned to risk assessments as a tool to make more accurate predictions of risk: specifically, the likelihood that a defendant, if released, would fail to return to court for their trial or would commit any crimes. The higher a defendant's risk, the more likely that a court is to detain that person until their trial. Pretrial detention is associated with a range of negative outcomes for the subject that include longer prison sentences, sexual abuse, and limited employment opportunities (see Chapter 8). Pretrial hearings are typically completed quickly, often within a matter of minutes [23]. Although pretrial decisions depend in part on the goal of ensuring that defendants return to court for trial without threatening public safety, they are also made with an interest in also protecting the liberty of defendants, ensuring that defendants are able to mount a proper defense, and reducing the hardship to defendants and their families [12]. Here, the "subject" is the criminal defendant and the "negative decision" is the decision to detain the defendant before trial (rather than release them).

Financial Loans

When someone applies for a financial loan, it is common for the potential lender to assess the risk that the borrower will fail to pay back the money (known as "defaulting" on the loan). This is often done using risk assessments that make predictions about the likelihood of loan default. The higher the risk that someone will default on the loan, the less likely the lender is to provide money to that person. Here, the "subject" is the loan applicant and the "negative decision" is the decision to reject the loan application (rather than approve the application).

A common type of loan provided by the U.S. government is home improvement loans (e.g., to rehabilitate a home or to make a home energy-efficient). In order to support low-income applicants who are unable to obtain affordable loans from banks, the government provides many types of home improvement loans [127]. These loans are motivated by a desire to promote equity, economic development, and community stability. Although there are no known cases of governments using risk assessments when giving out home improvement loans, this setting is akin to other government applications of risk assessments to determine whom should receive resources such as public benefits and housing [154, 418].

2.4.2 Data and Risk Assessments

We began our studies by creating risk assessments for pretrial detention and financial lending. In both settings, we used a dataset of historical cases to develop a risk assessment in the form of a machine learning classifier that predicted the probability of cases resulting in adverse outcomes. Our goal in this stage was not to develop optimal risk assessments, but to develop risk assessments that resemble those used in practice and that could be presented to participants during the Mechanical Turk experiments. The primary benchmark for the algorithms was that they make predictions more accurately than humans; given that that algorithms are more accurate than humans across a wide variety of tasks, this benchmark was essential for creating a realistic experimental environment.

Pretrial Data and Risk Assessment

To create our pretrial risk assessment, we used the dataset "State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties," which was collected by the U.S. Department of Justice [488]. The dataset contains court processing information pertaining to 151,461 felony cases that were filed during the month of May in even years from 1990-2006 and in 2009 in 40 of the 75 most populous counties in the United States. The data includes information about each case that includes the arrest charges, the defendant's demographic characteristics and criminal history, and the outcomes of the case related to pretrial release (whether the defendant was released

before trial and, if so, whether they were rearrested before trial or failed to appear in court for trial).

We first cleaned the dataset to prepare it for use developing a risk assessment. We cleaned the dataset to remove incomplete entries and restricted our analysis to defendants who were at least 18 years old, whose race was recorded as either Black or white. In order to have ground truth data about whether a defendant actually was rearrested before trial or failed to appear for trial, we also restricted our analysis to defendants who were released before trial.

This yielded a dataset of 47,141 defendants (Table 2.1). The defendants were primarily male (76.7%) and Black (55.7%), with an average age of 30.8 years. Among these defendants (all of whom were released before trial), 15.0% were rearrested before trial, 20.3% failed to appear for trial, and 29.8% exhibited at least one of these outcomes (which we defined as violating the terms of pretrial release).

	All N=47,141	**Black** N=26,246	**White** N=20,895
Background			
Male	76.7%	77.3%	75.5%
Black	55.7%	100.0%	0.0%
Mean age at arrest	30.8	30.1	31.8
Drug crime	36.9%	39.2%	34.0%
Property crime	32.7%	30.7%	35.3%
Violent crime	20.4%	20.9%	19.8%
Public order crime	10.0%	9.3%	10.8%
Has prior arrests?	63.4%	68.4%	57.0%
Mean number of prior arrests	3.8	4.3	3.1
Has prior convictions?	46.5%	51.2%	40.7%
Mean number of prior convictions	1.9	2.2	1.6
Has prior failure to appear?	25.1%	28.8	20.4%
Outcomes			
Rearrest	15.0%	16.9%	12.6%
Failure to appear	20.3%	22.6%	17.5%
Violation	29.8%	33.1%	25.6%

Table 2.1: Summary statistics for all of the defendants who were released before trial, broken down by defendant race. A violation means that the defendant was rearrested before trial, failed to appear for trial, or both.

We then used this data to train machine learning classifiers (i.e., our risk assessments) to predict the probability

that defendants would violate pretrial release (i.e., which defendants would be rearrested before trial or fail to appear in court for trial). We used the same method but generated new risk assessments for each study. We trained the models using gradient boosted decision trees [173] with the `xgboost` implementation in R [79]. The classifiers incorporated five features about each defendant: age, offense type, number of prior arrests, whether that person has any prior failures to appear, and number of prior convictions. Despite knowing the race and gender of defendants, we excluded these attributes from the models to match common practice among risk assessment developers [494]. Because our experiment participants would be predicting risk in increments of 10%, we rounded each risk assessment prediction to the nearest 10%.

The models achieved AUCs between 0.66-0.67, This indicates comparable accuracy to COMPAS [289, 243], the Public Safety Assessment [123], and other risk assessments [125]. According to a recent meta-analysis of risk assessments, our models have "Good" predictive validity [126]. We also evaluated the risk assessments for fairness and found that they are well calibrated. Given these evaluations, our pretrial risk assessments resemble those used within U.S. courts.

We selected from a sample of 300-500 defendants whose profiles would be shown to participants during the Mechanical Turk experiments. To protect defendant privacy, we could present to Turk participants information about only those defendants whose displayed attributes were shared with at least two other defendants in the full dataset. Although this restriction meant that we could not select a uniform random sample from the validation set, we found in practice that sampling from the validation set with weights based on each defendant's risk score yielded sample populations that resembled the full set of released defendants across most dimensions.

Financial Loans Data and Risk Assessment

To create our loans risk assessment, we used a dataset of loans from the peer-to-peer lending company Lending Club, which posts anonymized loan data on its website. The data contains records about all 2,004,091 loans that were issued between 2007 and 2018. Each record includes information such as the purpose of the loan; the loan

applicant's job, annual income, and approximate credit score; the loan amount and interest rate; and whether the loan was paid off. The data includes the first three digits each borrower's zip code but does not include further demographic information about loan applicants such as their age, race, or gender.

We cleaned the dataset to remove incomplete entries and classified credit scores into one of five categories (Poor, Fair, Good, Very Good, and Exceptional) defined by FICO [355]. We also limited the data to loans that have been either fully paid or defaulted on (although the data represents these loans as being "charged off," which is more extreme than defaulting on a loan, we refer to charged off loans as being defaulted on because the latter is the more commonly used and understood term).

As above, we used this data to train a risk assessment that could predict the probability that each loan would be defaulted on. Using cross-validation to find parameters and evaluate the models, we trained the classifier using gradient boosted decision trees [173] with the `xgboost` implementation in R [79]. Our model considered seven factors about each loan: the applicant's annual income, credit score category, and home ownership, as well as the loan's value, interest rate, monthly installment, and term of repayment (either 36 or 60 months). These models attained AUCs between 0.69-0.71, is similar to the performance of other loan default risk assessments that have been developed [493].

We selected uniform random samples of 300 loans that would be presented to the participants in each of our Mechanical Turk experiments. Unlike in the pretrial setting, there were no restrictions on which applicants we could present to participants.

2.4.3 Experiment Setup

In the second stage of each study, we conducted behavioral experiments on Amazon Mechanical Turk—a widely used online platform for human subjects research [59, 97]—to evaluate how the presentation of a risk assessment affects people's predictions and decisions.

The first study only included the loans setting. For the second and third studies, participants were sorted

into either the pretrial or loans setting and saw information only pertinent to that setting for the duration of the experiment. Participants were presented with a description of their setting in the tutorial. These descriptions explained the context of each setting, focusing on the key decision in each: for pretrial, whether to release or detain a criminal defendant before trial; for loans, whether to approve or reject a loan application. These descriptions also provided a brief discussion of the considerations that factor into these decisions: for pretrial, preventing flight risk and crime, but also ensuring the freedom of defendants who have not been proven guilty and preventing the harmful consequences of pretrial detention (such as losing one's job); for loans, preventing defaults but also enabling low-income homeowners to access resources, supporting economic development, and promoting neighborhood stability.

Each trial consisted of a consent page, a tutorial describing the task, an intro survey (to obtain demographic information and other participant attributes), the primary experimental task comprising a series of decisions or predictions, and an exit survey (to obtain participant beliefs and reflections on the task). The intro and exit surveys both included a simple question designed to ensure that participants were paying attention. We also included a comprehension test with several multiple-choice questions about the experiment at the end of the tutorial. Participants were not permitted to enter the study until they correctly answered all of the questions in the comprehension test. We restricted the task to Mechanical Turk workers inside the United States who had an historical acceptance rate of at least 75%. Turk workers were not allowed to participate in the experiment multiple times.

In both settings, participants were presented with narrative profiles about a random sample of subjects drawn from the 300- or 500-person populations drawn from the datasets of defendants and loan applicants. Profiles in the crime setting included the five features that the risk assessment incorporated as well as the race and gender of each defendant (we included these latter two features in the profiles despite not including them in the risk assessment because judges are exposed to these attributes in practice). Profiles in the loans setting included the same seven features that were included in the model. So that participants could look up background information

and the definitions of key terms when evaluating subjects, the information that had been presented during the tutorial (the description of the task and a glossary of key terms) was visible beneath these profiles. We used the same set of defendants and applicants for all treatments within a given experiment, allowing us to directly measure the impact of each treatment on the predictions and decisions made about each subject.

The primary research question of the study was to determine the effect of presenting risk assessment predictions about each defendant or applicant on the decisions made by participants, and in particular to determine whether presenting the risk assessment causes participants to weigh "risk" as a more salient factor in their decisions. Half of participants were therefore presented with the prediction made by the risk assessment about each subject, in addition to the narrative profiles about each subject, which were presented to every participant (for participants who would be shown the risk assessment, we included a description of the risk assessment in the tutorial and a question about the risk assessment in the comprehension test at the end of the tutorial). In other words, *the control group was participants making decisions without the risk assessment and the treatment group was participants making decisions with the risk assessment.*

The primary task for participants was to make predictions of risk for each defendant or applicant, on a scale from 0% to 100% in intervals of 10%. Participants in the pretrial setting were required to estimate the likelihood that criminal defendants will be arrested before trial or fail to appear in court for trial. Participants in the loans setting were required to estimate the likelihood that a loan applicant will default on their loan. In the first two studies, participants were asked to make only predictions; the third study had participants make both predictions and decisions.

All participants were paid a base sum for completing the study ($2 in the first two experiments, $3 in the third experiment). Participants making predictions also received an additional reward based on the accuracy of their predictions ($2 in the first two experiments, $1 in the third experiment). We allocated rewards following an inverted Brier score function: $\textit{score} = 1 - (\textit{prediction} - \textit{outcome})^2$, where $\textit{prediction} \in \{0,\ 0.1, \ldots,\ 1\}$ and $\textit{outcome} \in \{0,\ 1\}$ (because the sample populations were restricted to defendants who were released before trial

and loans that were granted, we have ground truth data about the binary outcome of each case). This inverted Brier score is bounded between 0 (worst possible performance) and 1 (best possible performance) and measures the accuracy of predictions about a binary outcome. We mapped the Brier score for each prediction to a payment such that perfect accuracy on all 40 predictions would yield the maximum bonus for that experiment. Because the Brier score is a proper score function [180], participants were incentivized to report their true estimates of risk. We articulated this to prediction-making participants during the tutorial and included a question about the reward structure in the comprehension test to ensure that they understood. We also measured false positive rates (using a threshold of 50%).

2.5 Limitations

A significant limitation of the experiments presented here is that our findings are based on the behaviors of Mechanical Turk workers rather than judges or loan agents, meaning that we cannot assume that the observed behaviors arise in practice. There are surely important divergences between how laypeople and experts respond to algorithms, particularly as it relates to trust and professional identity [55]. There are indications that our results accord with real-world outcomes, however: judges suffer from many of same cognitive illusions as other people [207], are skeptical about the benefits of algorithms [84, 77], and exhibit disparate interactions when using risk assessments [8, 107]. Indeed, many of the results presented in the following chapters align closely with behaviors and outcomes that have been observed in empirical studies of judges using risk assessments Continued research regarding the use of risk assessments in practice (and the relationship between behaviors observed in experimental versus natural settings) will provide vital evidence to inform ongoing debates about what role algorithms can or should play in consequential decisions.

Our studies also fail to capture the level of racial priming that could influence judges' use of risk assessments. While our experiments tell participants that a defendant is Black or white, a judge would also see the defendant's

name and physical appearance. Studies have shown that employers discriminate based on racially-indicative names [37] and that judges are harsher toward defendants with darker skin and more Afrocentric features [145, 272]. Thus, it is possible that the "disparate interactions" we observe in our experiments could be heightened in the practice, where race is more salient. Future research should study how people respond to risk assessments as racial priming increases.

The short length of each trial (25 predictions over approximately 20 minutes) means that we could not capture how the relationships between people and risk assessments evolve over extended periods of time. This is an important factor to consider when deploying algorithmic systems, especially given research demonstrating that the changes instigated by risk assessments are short-lived [464]. The immediate impacts of introducing algorithms into decision-making processes may not indicate the long-term implications of doing so.

A key aspect of future work will be to study algorithm-in-the-loop decision-making in real-world rather than experimental contexts. Mechanical Turk experiments are no substitute for *in situ* evaluations. However, experiments such as these provide an effective tool for diagnosing the types of human-algorithm interactions that could arise in practice. Issues identified in experiments can inform the design and evaluation of real-world systems in order to prevent breakdowns when the stakes are high.

Chapter 3

Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments

3.1 Introduction

This study sheds new light on how risk assessments influence human decisions in the context of pretrial adjudication. We ran experiments using Amazon Mechanical Turk to study how people make predictions about risk, both with and without the aid of a risk assessment. We focus on pretrial release, which in many respects resembles a typical prediction problem.[1] By studying behavior in this controlled environment, we discerned important patterns in how risk assessments influence human judgments of risk. Although these experiments involved laypeople rather than judges—limiting the extent to which our results can be assumed to directly im-

[1] After someone is arrested, courts must decide whether to release that person until their trial. This is typically done by setting an amount of "bail," or money that the defendant must pay as collateral for release. The broad goal of this process is to protect individual liberty while also ensuring that the defendant appears in court for trial and does not commit any crimes while released (whether the defendant is guilty of the offense that led to the arrest is not a factor at this stage). In order to make pretrial release decisions, judges must determine the likelihood—or the "risk"—that the defendant, if released, will fail to appear in court or will be arrested.

plicate real-world risk assessments—they highlight several types of interactions that should be studied further before risk assessments can be responsibly deployed in the courtroom.

Before running our experiments, we made three hypotheses:

Hypothesis 1 (Performance). Participants presented with a risk assessment will make predictions that are less accurate than the risk assessment's.

Hypothesis 2 (Evaluation). Participants will be unable to accurately evaluate their own and the algorithm's performance.

Hypothesis 3 (Bias). As they interact with the risk assessment, participants will be disproportionately likely to increase risk predictions about Black defendants and to decrease risk predictions about white defendants.

Our results suggest several ways in which the interactions between people and risk assessments can generate errors and biases in pretrial predictions, thus calling into question the supposed efficacy and fairness of risk assessments. First, even when presented with the risk assessment's predictions, participants made decisions that were less accurate than the advice provided. Second, people could not effectively evaluate the accuracy of their own or the risk assessment's predictions: participants' confidence in their performance was *negatively* associated with their actual performance and their judgments of the risk assessment's accuracy and fairness had no association with the risk assessment's actual accuracy and fairness. Finally, participant interactions with the risk assessment introduced two new forms of bias (which we collectively term "disparate interactions") into decision-making: when evaluating Black defendants, participants were 25.9% more strongly influenced to increase their risk prediction at the suggestion of the risk assessment and were 36.4% more likely to deviate from the risk assessment toward higher levels of risk. Further research is necessary to ascertain whether judges exhibit similar behaviors.

3.2 Methods

3.2.1 Study Design

See Section 2.4 for the full details of the study design. Here, I describe the elements of this experiment that are particular to this study. This study included only the pretrial setting and did not include the loans setting.

Data and Risk Assessments

After developing the risk assessment, we evaluated it for fairness. As Figure 3.1 indicates, the model is well-calibrated: at every risk score from 10% to 60% (the full range of risks predicted), Black and white defendants are statistically equally likely to violate pretrial release. We focused on calibration not as an ideal metric for fairness (recognizing that no perfect metric for fairness can exist [200]), but because it is the most commonly-used approach for evaluating risk assessments in practice [278, 166, 130]. In fact, similarly to COMPAS [17], we find that our model disproportionately makes false positive errors for Black defendants compared to white defendants (7.0% versus 4.6%, assuming a naïve threshold of 50%).

For this study, we selected an experimental sample of 500 defendants whose profiles would be presented to both the control and treatment groups during the experiments (Table 3.1).

Experiment Setup

Each participant was presented with narrative profiles about a random sample of 25 defendants drawn from the 500-person experiment sample population, and was asked to predict each defendant's risk (Figure 3.2).

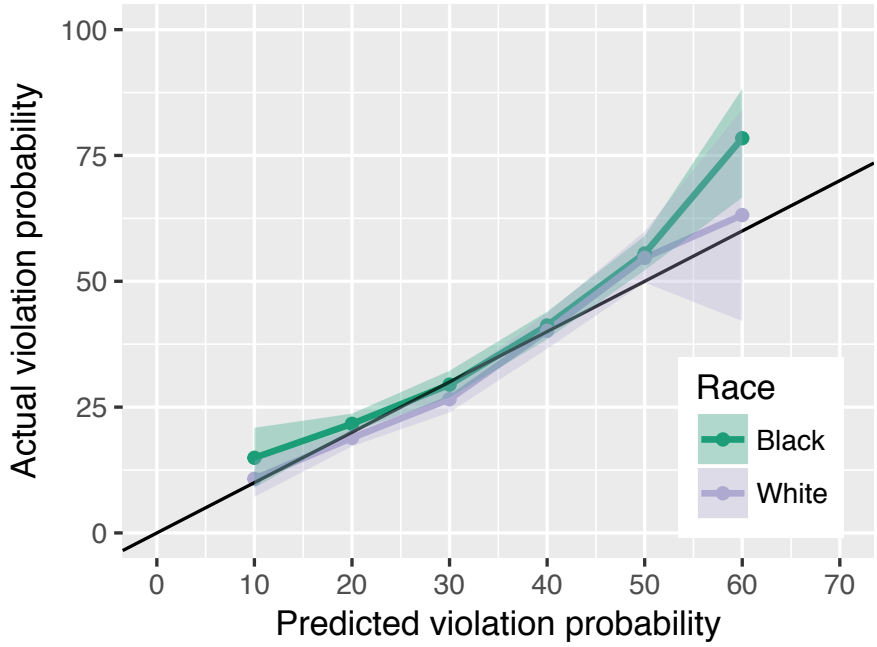

Figure 3.1: Comparison of risk assessment predictions and actual violation probabilities for Black and white defendants, indicating that the model is well-calibrated across race. Bands indicate 95% confidence intervals.

	Experiment Sample N=500	**Sample (Black)** N=303	**Sample (White)** N=197
Background			
Male	79.6%	81.9%	76.1%
Black	60.6%	100.0%	0.0%
Mean age	28.7	27.7	30.3
Drug crime	42.6%	44.9%	39.1%
Property crime	34.6%	33.3%	36.5%
Violent crime	17.8%	17.8%	17.8%
Public order crime	5.0%	4.0%	6.6%
Prior arrest(s)	54.4%	61.7%	43.1%
# of prior arrests	3.5	4.2	2.4
Prior conviction(s)	35.4%	40.9%	26.9%
# of prior convictions	2.0	2.3	1.5
Prior failure to appear	27.6%	33.0%	19.3%
Outcomes			
Rearrest	15.4%	17.8%	11.7%
Failure to appear	19.6%	19.5%	19.8%
Violation	29.8%	31.4%	27.4%

Table 3.1: Summary statistics for the 500 defendants presented to participants. See Table 2.1 for the attributes of the full data sample.

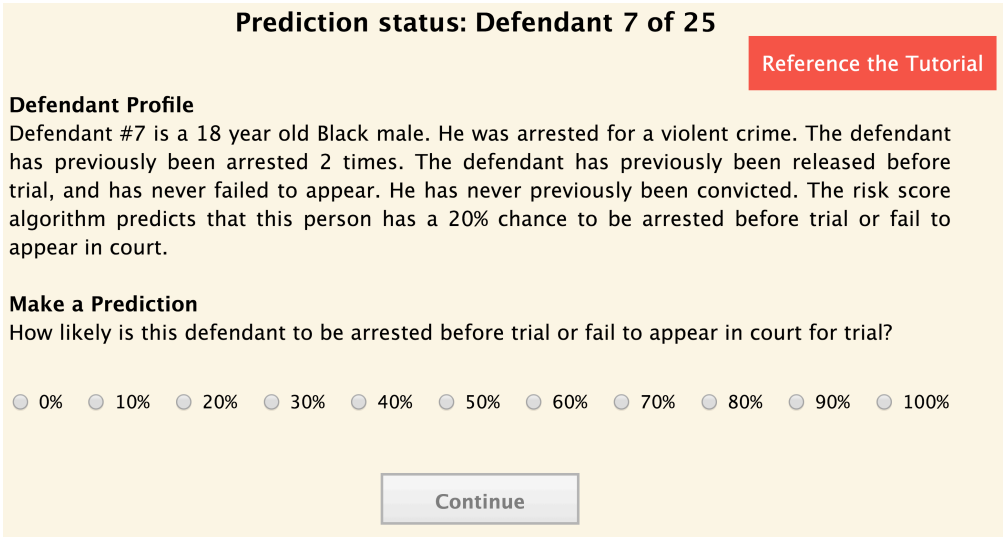

Figure 3.2: An example of the prompt presented to participants in the treatment group. Participants in the control group saw the same prompt, but without the sentence about the risk score algorithm.

3.2.2 Analysis

Because we presented the same set of 500 defendants to both the control and treatment groups, we could measure the influence of the risk scores on the predictions about each defendant by comparing the predictions made by the control and treatment groups. For each defendant j, we defined the risk score's influence

$$I_j = \frac{t_j - c_j}{r_j - c_j} \tag{3.1}$$

where t_j and c_j are the average predictions made about that defendant by participants in the treatment and control groups, respectively, and r_j is the prediction made by the risk assessment. An $I = 0$ means that, on average, the treatment group makes identical predictions to the control group, completely discounting the risk score, while an $I = 1$ means that the treatment group makes identical predictions to the risk score.[2] This measure of influence is similar to the "weight of advice" metric that has been used to measure how much people alter their decisions

[2] Although I will mostly fall between 0 and 1, it is possible for I to fall outside these bounds if participants move in the opposite direction than the risk assessment suggests or adjust beyond the risk assessment.

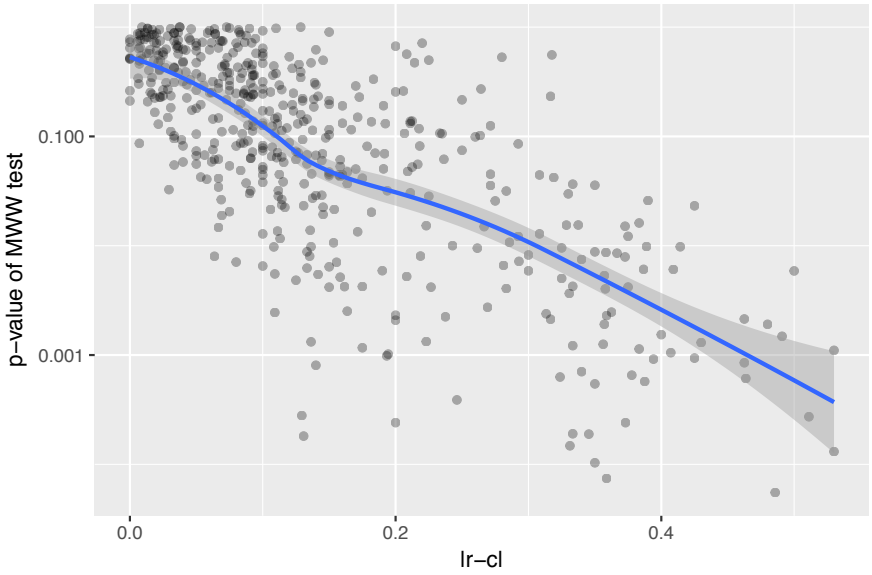

Figure 3.3: Comparison of x) the difference between the risk score r and the control group's average prediction c and y) the difference between the distributions of predictions made by the control and treatment groups, as measured by the p-value of a Mann-Whitney-Wilcoxon (MWW) test. Each dot represents one defendant and is made partially transparent such that darker regions represent clusters of data. The blue line and gray band represent a local regression (LOESS) smoothing fit and 95CI. As r and c diverge, the treatment and control group prediction distributions also diverge. This indicates that, although our analyses focused on the average predictions made by the control and treatment groups, the risk assessment influenced the full distribution of predictions made by the treatment group.

when presented with advice [519, 309]. Comparing the distributions of predictions made by the control and treatment groups indicates that the risk assessment influences the full distribution of predictions made by the treatment group, not just the average (Figure 3.3). To obtain reliable measurements, when evaluating algorithm influence we excluded all predictions about the 112 defendants for whom $|r_j - c_j| < 0.05$.

We used a variant of Equation 3.1 to measure the influence of the risk assessment on each participant in the treatment group. For every prediction made by a participant, we measured the risk assessment's influence by taking that prediction in place of the average treatment group prediction. We then averaged these influences across the 25 predictions that the participant made. That is, the influence of the risk assessment on participant k is

$$I^k = \frac{1}{25} \sum_{i=1}^{25} \frac{p_i^k - c_i}{r_i - c_i} \tag{3.2}$$

where p_i^k refers to participant k's prediction about the ith defendant (out of 25) presented.

Our primary dimension of analysis was to compare behavior and performance across the race of defendants, which has been at the crux of debates about fairness in pretrial and sentencing risk assessments [17, 82**?**]. Similar audits should be conducted across other intersecting forms of identity, such as gender and class [108].

3.3 Results

We conducted trials on Mechanical Turk over the course of a week in June 2018 (in 6 batches over 4 weekdays and 2 weekend days, at times ranging from morning to evening to account for variations in the population of Turk workers). 601 workers completed the experiment; we excluded all data from participants who failed at least one of the attention check questions or who required more than three attempts to pass the comprehension test. This process yielded a population of 554 participants (Table 3.2). The participants were 58.5% male and 80.5% white, and the majority (65.5%) have completed at least a college degree. We asked participants to self-report their familiarity with machine learning and the U.S. criminal justice system on a scale from 1 ("Not at all") to 5 ("Extremely").

During the exit surveys, participants reported that the experiment paid well, was clear, and was enjoyable. Participants earned an average bonus of $1.54 (median=$1.56), making the average total payment $3.54. Participants completed the task in an average of 20 minutes (median=12), and earned an average wage of $20 per hour (median=$18). Out of 213 participants who responded to a free text question in the exit survey asking for any further comments, 32% mentioned that the experiment length and payment were fair. Participants were also asked in the exit survey to rate how clear and enjoyable the experiment was, on a scale from 1 to 5. The average rating for clarity was 4.4 (55% of participants rated the experiment clarity a 5), and the average rating for enjoyment was 3.6 (56% rated the experiment enjoyment a 4 or 5).

The participants cumulatively made 13,850 predictions about defendants, providing us with 13.85 ± 3.9 predictions about each defendant's risk under each of the two experimental conditions.

	All N=554	Control N=250	Treatment N=304
Male	58.5%	60.4%	56.9%
Black	7.4%	8.0%	6.9%
White	80.5%	80.0%	80.9%
18-24 years old	9.7%	7.6%	11.5%
25-34 years old	42.4%	43.6%	41.4%
35-59 years old	43.9%	44.4%	43.4%
60-74 years old	4.0%	4.4%	3.6%
College degree or higher	65.5%	67.6%	63.8%
Criminal justice familiarity	2.8	2.9	2.8
Machine learning familiarity	2.4	2.3	2.4
Experiment clarity	4.4	4.5	4.4
Experiment enjoyment	3.6	3.6	3.7

Table 3.2: Attributes of the participants in our experiments.

	Control N=6,250	Treatment N=7,600	Risk assessment N=7,600
Average reward	0.756	0.786	0.807
False positive rate	17.7%	14.8%	10.1%

Table 3.3: The first two columns show the performance of participants within the control and treatment groups and the third column shows the performance of the risk assessment (N is the total number of predictions made). Two-sided t-tests and χ^2 tests confirm that the average rewards and the false positive rates, respectively, of all three prediction approaches are statistically distinct from one another (all with $p < 10^{-5}$).

3.3.1 Hypothesis 1 (Performance)

Participants in the treatment group earned a 4.0% larger average reward and a 16.4% lower false positive rate than participants in the control group (Table 3.3). A two-sided t-test and χ^2 test confirm that these differences are statistically significant (both with $p < 10^{-5}$). A regression of each participant's performance on their treatment and personal characteristics found that being in the treatment group was associated with a 0.03 higher average reward ($p < 10^{-7}$). The only personal attribute that had a significant relationship with average reward was gender (women performed slightly better than men, with $p = 0.045$).

Yet although presenting the risk assessment improved the performance of participants, the treatment group

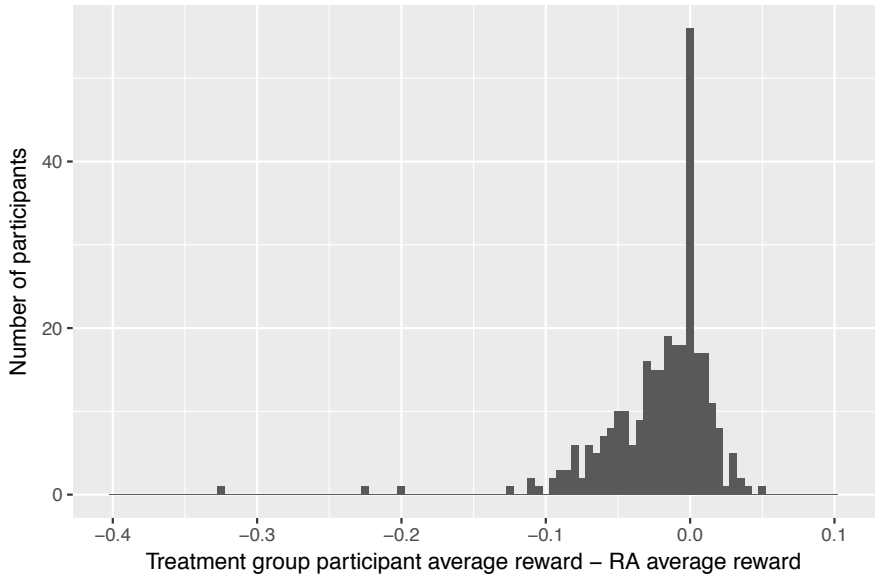

Figure 3.4: Distribution of differences between participant performance and risk assessment (RA) performance over the course of each treatment group participant's trial. Negative values indicate that the treatment group participant received a lower average reward than the risk assessment for the 25 predictions that the participant made. Out of the 304 treatment group participants, 195 (64.1%) earned a lower average reward than the risk assessment, 37 (12.2%) earned an equal average reward, and 72 (23.7%) earned a larger average reward.

significantly underperformed the the risk assessment (Table 3.3). Despite being presented with the risk assessment's predictions, the treatment group achieved a 2.6% lower average reward and a 46.5% higher false positive rate than the risk assessment (both with $p < 10^{-8}$). Only 23.7% of participants in the treatment group earned a higher average reward than the risk assessment over the course of their trial, compared to 64.1% who earned a lower reward than the risk assessment (Figure 3.4).

We broke these results down by race to compare how participants and the risk assessment performed when making predictions about Black and white defendants. As Figure 3.5 indicates, a similar pattern was true for both races: the treatment group outperformed the control group but underperformed the risk assessment. Taking the control group performance as a lower bound and the risk assessment performance as an upper bound, the treatment group achieved a similar relative improvement in its predictions about both races: for average reward, 58.7% of possible improvement for Black defendants and 59.7% for white defendants; for false positive rate, 39.3% of possible improvement for Black defendants and 39.5% for white defendants (neither difference across

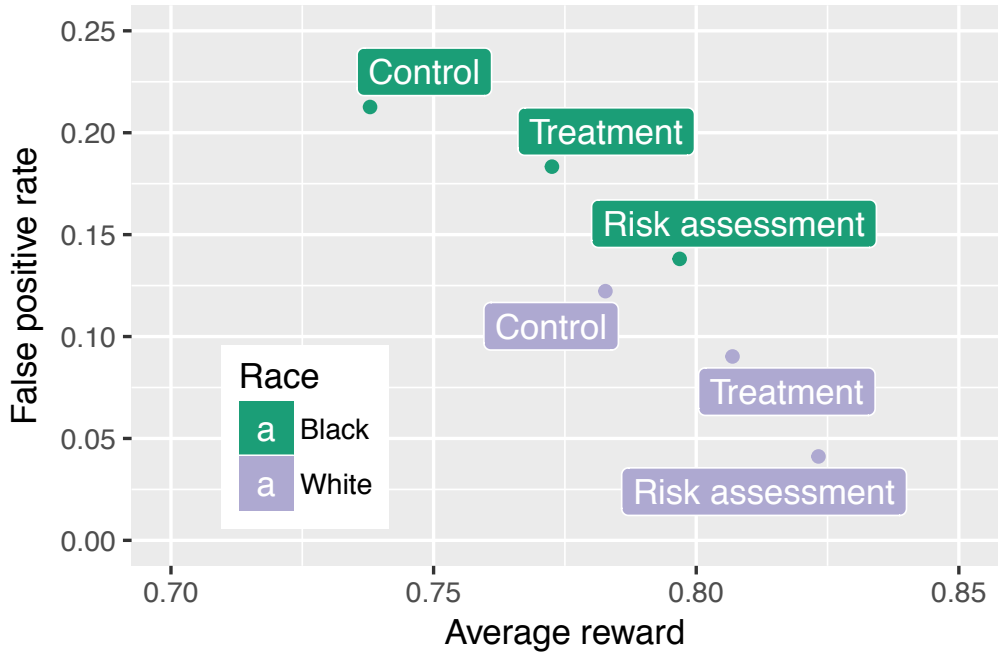

Figure 3.5: Performance of the control group, treatment group, and risk assessment, broken down by defendant race. In both cases, the treatment group outperforms the control group but underperforms the risk assessment.

race is statistically significant).

The actual performance level differs significantly across race, however. All three prediction approaches (i.e., the control group, the treatment group, and the risk assessment) achieve a larger reward and lower false positive rate for white defendants than for Black defendants (all with $p < 10^{-6}$). Most notably, the treatment group attains a 4.5% higher average reward for white than Black defendants and its false positive rate for Black defendants (18.3%) is more than double its false positive rate for white defendants (9.0%).

3.3.2 Hypothesis 2 (Evaluation)

To assess whether participants could evaluate the quality of their predictions, we compared their self-reported confidence (from the exit survey) to their actual performance, as measured by their average Brier reward during the task. The average participant confidence was 3.2 (on a scale from 1 to 5), with the reward decreasing as reported confidence increases (Figure 3.6). We regressed confidence on performance (controlling for each participant's treatment, demographic information, and exit survey responses) and found that average reward

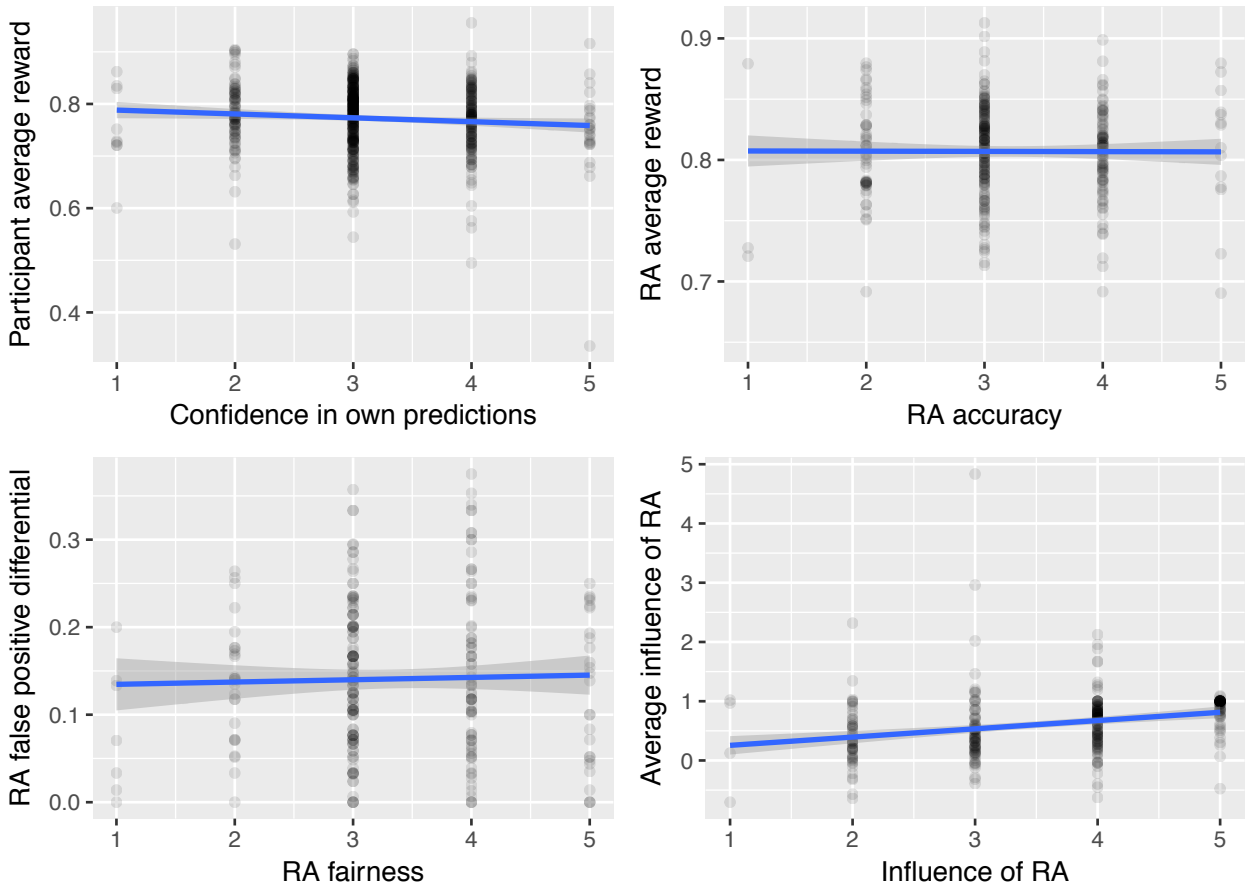

Figure 3.6: Comparison of participant evaluations and the actual behaviors of themselves and the risk assessment (RA). Each x-axis represents the participant reflection provided for the first four questions of the exit survey; the y-axes represent a proxy for the actual outcome that the participant was evaluating (as described in Section 3.3.2). Each dot represents one participant and is made partially transparent such that darker regions represent clusters of data. The linear regression fits presented here do not include the controls described in Section 3.3.2, but are shown for demonstration purposes, as the fits depicted closely resemble the relationships found in the full regression analyses.

was negatively associated with confidence ($p = 0.0186$). In other words, the more confidence participants expressed in their predictions, the less well they actually performed. This pattern holds across both the control and treatment groups.

We next analyzed whether participants in the treatment group could evaluate the risk assessment's accuracy, as measured by its average Brier reward on the 25 defendants presented to the participant (these average rewards ranged from 0.69 to 0.91). We regressed the participants' evaluations of the risk assessment's accuracy against the risk assessment's actual performance, while controlling for each participant's performance, demographic

information, and exit survey responses. The participant's evaluation of the risk assessment's accuracy did not have any significant relationship with the risk assessment's performance during the task, suggesting that participants were unable to perceive any differences in risk assessment accuracy over the samples they observed (Figure 3.6).

We also considered whether participants could discern how fairly the risk assessment made predictions. As a rough measure of algorithmic fairness during each trial, we measured the difference between the risk assessment's false positive rates for Black and white defendants on the 25 defendants presented to the participant (in order to focus on the most salient aspect of bias, we restricted this analysis to the 81% of participants for whom the risk assessment had a greater or equal false positive rate for Black than white defendants). Regressing participant evaluations of the risk assessment's fairness on the risk assessment's false positive rate differences (controlling for each participant's performance, demographic information, and exit survey responses, along with the risk assessment's performance) found no significant relationship between perceived and actual fairness (Figure 3.6).

Finally, we evaluated whether participants in the treatment group could recognize how heavily they incorporated the risk assessment into their decisions. Regressing the participants' self-reports of influence on the extent to which they were actually influenced by the risk assessment (using the risk score influence measure introduced in Equation 3.2, and controlling for each participant's performance, demographic information, and exit survey responses, along with the risk assessment's performance) indicates that participants could generally discern how strongly they were influenced by the risk assessment ($p < 10^{-4}$; Figure 3.6).

3.3.3 Hypothesis 3 (Bias)

We interrogated Hypothesis 3 through two complementary approaches: first, by taking the control group's predictions as the baseline participant predictions to measure the risk assessment's influence on the treatment group, and second, by taking the risk assessment's predictions as the starting point to measure how much and in which direction the treatment group participants deviated from those predictions.

Although we could not precisely discern how participants made decisions, the responses to an optional free

Deviated from the risk assessment (N=79, 50.6%; average reward=0.79)
"I used the risk scores as a starting point and then I made adjustments based on my own intuition about each case."
"I used them as an anchor point, and then shifted up or down one depending on my personal feelings about the individual cases."
Incorporated the risk assessment after making own judgment (N=58, 37.2%; average reward=0.79)
"I did not consider it until after making my own decision and then adjusted accordingly."
"decided on a score myself first, then I let the risk score slightly sway my decision."
Followed the risk assessment completely (N=10, 6.4%; average reward=0.81)
"I input exactly what the risk score indicated. It's probably smarter than I am."
"I used the risk score all the time for the entire HIT. Machine learning is much more accurate than humans."
Ignored the risk assessment entirely (N=9, 5.8%; average reward=0.77)
"I just went with my own thoughts after reading each scenario."
"I didn't really pay that much attention to it since I felt the percentages were too low."

Table 3.4: A representative sample of the responses that treatment group participants submitted when asked on the exit survey about how they incorporated the risk scores into their decisions, broken down by the general strategy they indicate having used.

response question in the exit survey about how participants used the risk scores suggest that people predominantly followed a mix of these two approaches. Out of the 156 participants who described their strategy, 79 (50.6%) used the risk assessment as a baseline, 58 (37.2%) made their own judgment and then incorporated the risk assessment, 10 (6.4%) followed the risk assessment completely, and 9 (5.8%) ignored the risk assessment entirely (Table 3.4). The group that followed the risk assessment earned the largest average reward (0.81), while the group that ignored the risk assessment earned the lowest (0.77). The other two groups both earned average rewards of 0.79, and were statistically indistinguishable.

Analyzing behavior through the lens of the two most common strategies yields complementary evidence for "disparate interactions," i.e., interactions with the risk assessment that lead participants to disproportionately make higher risk predictions about Black defendants and lower risk predictions about white defendants.

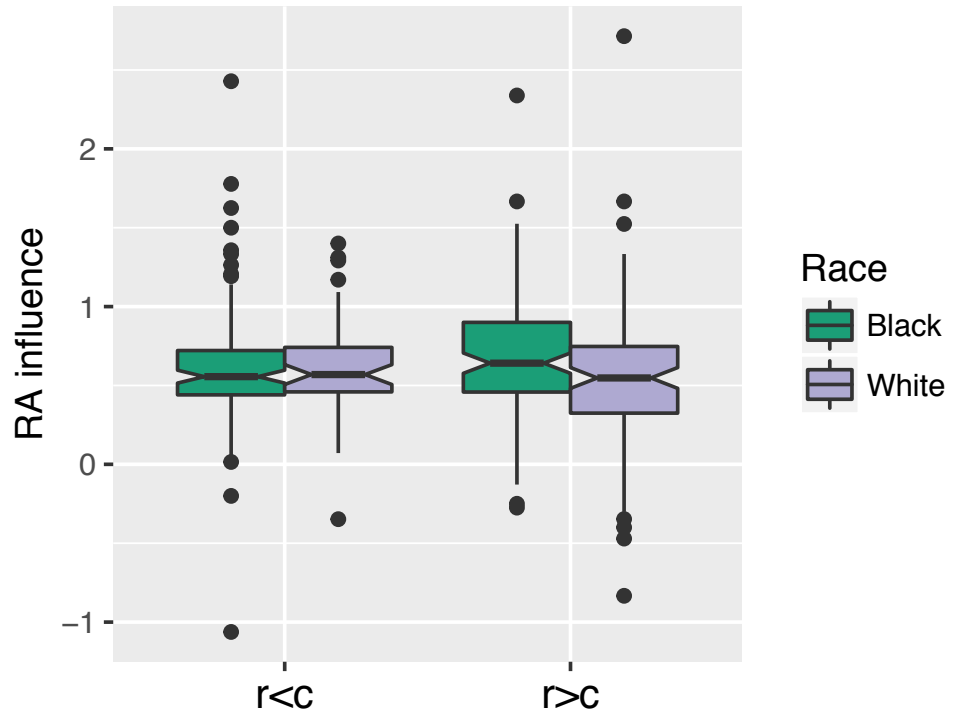

Figure 3.7: The influence of the risk assessment (RA) on participant predictions, broken down by whether the risk score is less or greater than the control group's average prediction ($r < c$ and $r > c$, respectively), and compared across the race of defendants. While the risk assessment's influence is nearly identical across race when $r < c$, when $r > c$ the risk assessment exerts a 25.9% stronger influence on participants who are evaluating Black defendants ($p = 0.02$).

Influence of risk scores

Because we presented the same population of defendants to the control and treatment groups, we could directly measure how presenting the risk score to participants affected the predictions made about each defendant. For each defendant, we measured the influence of the risk assessment on the treatment group's predictions as described in Equation 3.1 (excluding the 112 defendants for whom $|r_j - c_j| < 0.05$). The risk assessment exhibited an average influence of 0.61; as this number is greater than 0.5, it suggests that treatment group participants placed more weight on the risk assessment than on their own judgment. A two-sided t-test found no statistically significant difference between the risk assessment's influence when its prediction was less or greater than the control group's prediction ($r < c$ or $r > c$, respectively).

Splitting the defendants by race tells a more complex story (Figure 3.7). When the risk score was lower than

the control group's average prediction ($r < c$), the risk assessment exerted a similar influence on participants regardless of the defendant's race (0.61 vs. 0.60; p=0.77). Yet when the risk assessment predicted a higher risk than the control group ($r > c$), it exerted a 25.9% stronger average influence on predictions about Black defendants than on predictions about white defendants (0.68 vs. 0.54; a two-sided t-test finds $p = 0.02$ and 95CI of the difference in means [0.02, 0.25]).

This outcome cannot be explained by differences in the raw disparities between the risk assessment's and the control group's predictions (i.e., the value of $r - c$), since the values of $r - c$ do not differ significantly across defendant race (the average disparity for both races is 0.25 when $r < c$ and 0.11 when $r > c$). Breaking out Figure 3.7 based on the value of $r - c$ indicates that the risk assessment exerts an equal influence on predictions about both races at all values of $r - c$, except for when $r - c = 0.1$ (Figure 3.8).

Thus, the risk assessment leads to larger increases in risk for Black defendants (as measured by $t - c$). While the shift in participant predictions precipitated by the risk assessment is identical when $r < c$ (the risk assessment generates an average reduction of 0.14 for both Black and white defendants), when $r > c$ the average increase for Black defendants is 0.075 while the average increase for white defendants is 0.063. Although these results are not significant (a two-sided t-test finds $p = 0.076$ and 95CI difference in means [-0.001, 0.02]), considering each prediction from the treatment group independently, rather than taking averages for each defendant (i.e., replacing t_j with p_j^k in Equation 3.1), yields further evidence for this result: the average increase for Black defendants is 0.077 compared to 0.064 for white defendants (a 20.3% larger average increase), with $p = 0.003$ and 95CI difference in means [0.004, 0.02]. Moreover, among defendants for whom $r - c = 0.1$, the increase in participant risk prediction instigated by the risk assessment is 25.5% larger for Black defendants ($p = 0.042$; Figure 3.8).

We ran linear regressions to see what determines the risk assessment's influence on participants. We split defendants into two categories—those for whom $r < c$ (Group 1) and those for whom $r > c$ (Group 2). For each group, we regressed the algorithm's influence on predictions about each defendant (Equation 3.1) on that defendant's demographic attributes and criminal background, along with the value of $|r - c|$. For Group 1, the

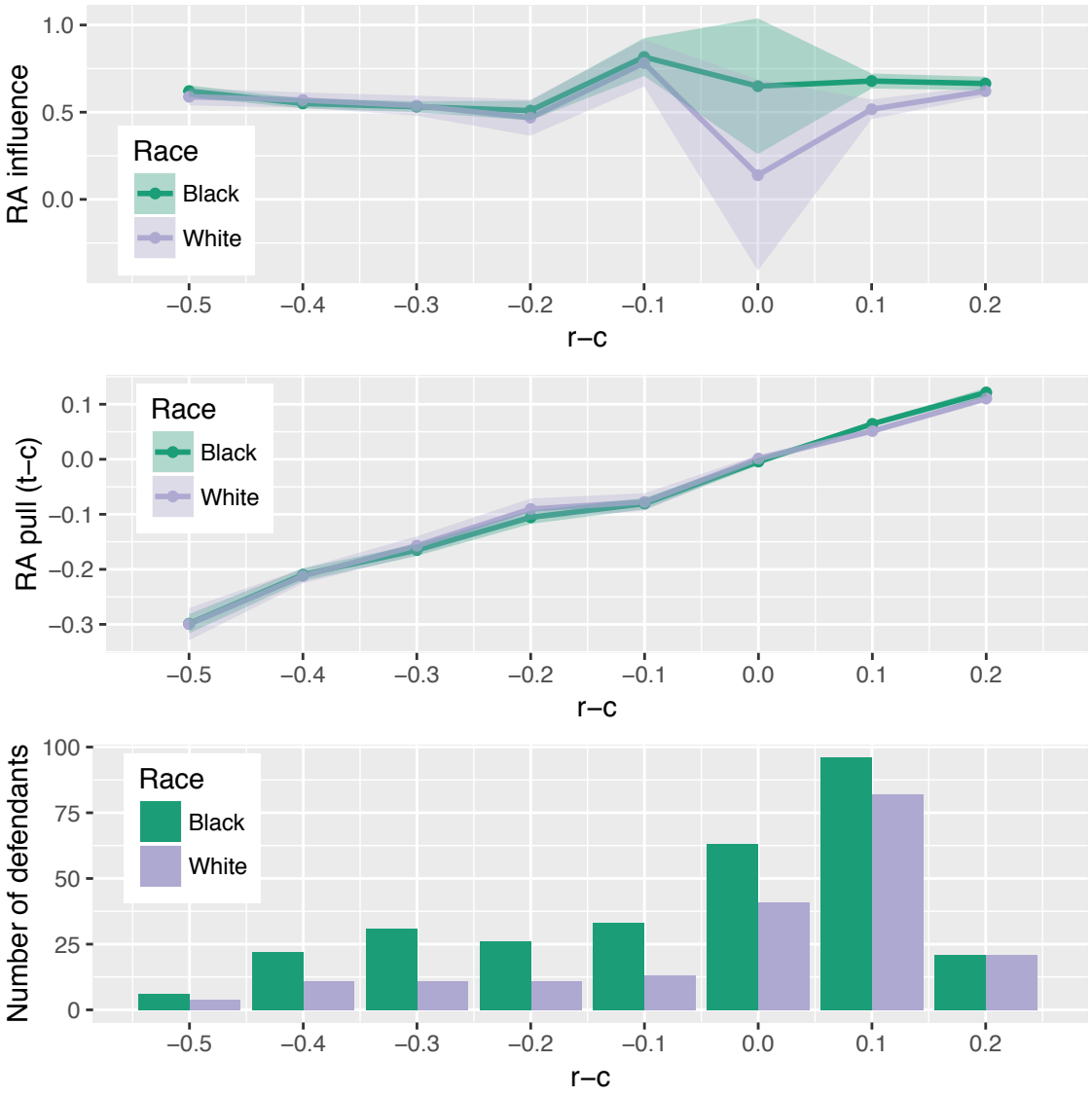

Figure 3.8: Top: The average influence of the risk assessment on treatment group participants (as measured by Equation 3.1) based on defendant race and the difference between the risk assessment and control group predictions $(r - c)$, rounded to the nearest 0.1. The bands depict the standard error for each group; the standard errors around the $r - c = 0$ groups are particularly large because (given that $r - c$ is the denominator of Equation 3.1) the influence measurements become unstable when r and c are almost identical (for this reason we excluded the eight defendants for whom $r = c$ from all three panels). The differences in the risk assessment's influence across race are statistically significant only when $r - c = 0.1$: the average influence on participants evaluating Black defendants is 0.68 while the average influence on participants evaluating white defendants is 0.52 ($p = 0.02$, 95CI difference in means [0.02,0.30]). Middle: The actual change in risk prediction instigated by the risk assessment (i.e., $t - c$, the numerator of Equation 3.1). The differences in the risk assessment's pull across race are statistically significant only when $r - c = 0.1$: the average increase for Black defendants is 0.064 while the average increase for white defendants is 0.051 ($p = 0.042$, 95CI difference in means [0.0005, 0.0255]). Bottom: The number of Black and white defendants who fall into each category.

risk assessment exerted more influence as $|r - c|$ increased, but less influence for defendants with a previous failure to appear on their records. For Group 2, the risk assessment similarly was more influential as $|r - c|$ increased. Three other attributes were also statistically significant: the risk assessment exerted more influence on participants making predictions about Black defendants, defendants who were arrested for a violent crime, and defendants with more prior convictions. Thus, when $r > c$, participants were more strongly influenced to increase their risk predictions for Black defendants in two ways: they responded both directly to race and to a feature that is correlated with race (prior convictions; Table 2.1).

Participant deviations from risk scores

For each prediction made by participants in the treatment group, we measured how far and in which direction that prediction deviated from the risk assessment's recommendation. That is, we measured $d_j^k = p_j^k - r_j$. The average deviation among the 7600 treatment group predictions was 0.014, with a median deviation of 0. Participants deviated to a higher risk prediction 26.9% of the time, matched the risk assessment 40.8% of the time, and deviated to a lower risk prediction 32.3% of the time. The results from Section 3.3.1 suggest that these deviations tend to make participant predictions less accurate than the risk assessment.

As in the previous section, these statistics differ by defendant race. While the average deviation for white defendants was -0.002, the average deviation for Black defendants was 0.024 ($p = 7 \times 10^{-13}$, 95CI difference in means [0.019, 0.033]). This difference emerged because participants were more likely to deviate positively from the risk assessment when evaluating Black defendants and to deviate negatively when evaluating white defendants (the average deviation magnitude was the same across race for both positive and negative deviations). As Figure 3.9 depicts, participants deviated to a higher risk prediction 30.0% of the time for Black defendants compared to 22.0% of the time for white defendants (36.4% more), and conversely deviated to a lower risk prediction 29.2% of the time for Black defendants compared to 37.2% of the time for white defendants (21.5% less). Participants matched the risk assessment in 40.8% of predictions when evaluating both races.

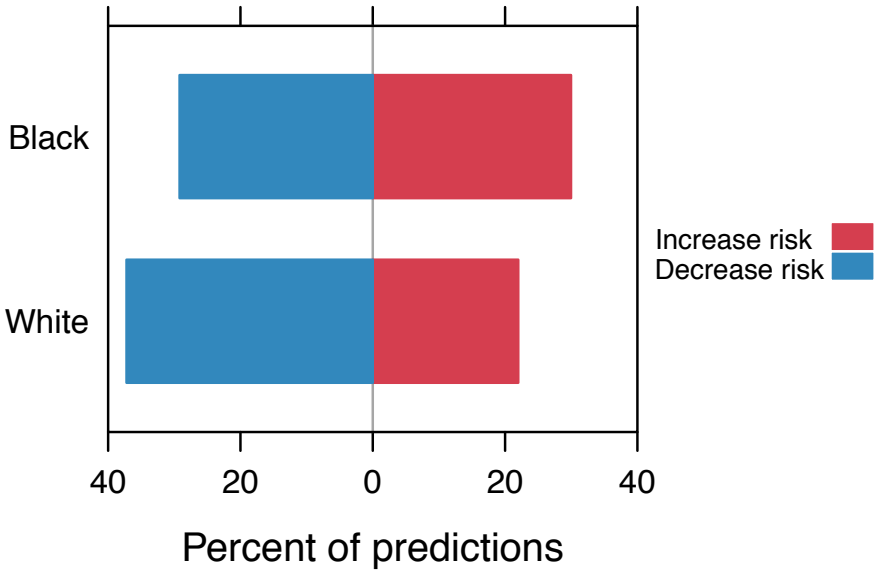

Figure 3.9: The rate at which participants deviated from the risk assessment's prediction toward higher and lower levels of risk, broken down by defendant race. When evaluating Black defendants, participants were 36.4% more likely to deviate positively from the risk assessment and 21.5% less likely to deviate negatively (participant predictions matched the risk assessment at an equal rate for both races).

We regressed each deviation on the characteristics of the defendant and the participant, the prediction made by the risk assessment, and the participant's status in the experiment (i.e., which in the sequence of 25 predictions the participant was making). Since these deviations include repeated samples for each defendant and participant, we used a linear mixed-effects model with random effects for the defendant and participant identities. Several characteristics of defendants had statistically significant associations with the deviations: participants were more likely to deviate positively from the risk assessment when evaluating younger defendants, defendants arrested for a violent crime, defendants with more prior arrests and convictions, and defendants with a prior failure to appear. Neither the defendant's race nor any attributes of participants had a statistically significant relationship with deviations.

These results suggest that while participants did not deviate from the risk assessment based explicitly on race, they deviated based on attributes that are unevenly distributed across race: compared to white defendants, Black defendants on average have more prior arrests, convictions, and failures to appear (Table 2.1).

3.4 Discussion

This study presents initial evidence regarding how risk assessments influence human decision-makers. Confirming our three hypotheses, our results indicate that people underperform risk assessments even when provided with its advice; are unable to evaluate the performance of themselves or the risk assessment; and engage in "disparate interactions," whereby their use of risk assessments leads to higher risk predictions about Black defendants and lower risk predictions about white defendants.

This work demonstrates how theoretical evaluations are necessary but insufficient to evaluate the impacts of risk assessments: what appears to be a fair source of information can, depending on how people interact with it, become a leverage point around which discrimination manifests. It is necessary to place risk assessments into a sociotechnical context so that their full impacts can be identified and evaluated.

Our results highlight a significant but often overlooked aspect of algorithmic decision-making aids: introducing risk assessments to pretrial decisions does not eliminate discretion to create "objective" judgments, as many have argued [217, 361, 115]. Instead, risk assessments merely shift discretion to different places, which include the judge's interpretation of the assessment and decision about how strongly to rely on it. This reality must become a central consideration of any proposals for and evaluations of risk assessments, especially given that previous attempts to standardize the criminal justice system—sentencing reform efforts in the 1980s—shifted discretion to prosecutors, generating a racially-biased rise in excessive punishment [315].

A particular danger of judicial discretion about how to incorporate risk assessments into decisions is the potential for disparate interactions: biases that emerge as an algorithmic prediction filters through a person into a decision. Our experiment participants were 25.9% more strongly influenced by the risk assessment to increase their risk prediction when evaluating Black defendants than white ones, leading to a 20.3% larger average increase for Black than white defendants due the risk assessment. Moreover, participants were 36.4% more likely to deviate positively from the risk assessment and 21.5% less likely to deviate negatively from the risk assessment

when evaluating Black defendants.[3]

These disparate interactions emerged through both direct and indirect bias: while race had a direct role in increasing the risk score's influence on participants, the disparities in influence and deviations also arose due to participants responding to particularly salient features that are unevenly distributed by race (such as number of prior convictions)—essentially double-counting features for which the risk assessment had already accounted. This behavior resembles that of machine learning algorithms, which can be racially biased even when race is not included as an explicit factor [17], and highlights the importance of studying the complex mechanisms through which discrimination can manifest. Future work should explore how different ways of presenting and explaining risk assessments (and of training people to use them) could improve performance and in particular reduce disparate interactions.

An important research direction that could guide such efforts is to study the processes through which people make decisions when provided with risk assessments. Our participants followed several approaches when evaluating defendants, the most common being using the risk assessment to influence their initial judgment and using the risk assessment as a baseline (Table 3.4). Analyzing participant behavior from both of these perspectives indicated related forms of disparate interactions. Meanwhile, the most successful strategy was to directly follow the risk assessment. While in theory it is possible for people to synthesize the risk assessment with their own judgment to make better decisions than either could alone, in practice we found no evidence that any strategy taken by participants leads them to outperform the risk assessment.

A major limitation to people's use of risk assessments is their inability to evaluate their own and the risk assessment's performance. Many proponents defend the deployment of risk assessments on the grounds that judges have the final say and can discern when to rely on the predictions provided [514, 243, 291]. But our results indicate that this is an unrealistic expectation: our participants' judgments about their own performance were

[3]Although it is possible that participants predicted higher risk for Black defendants to account for the racial bias in arrests, we do not believe this was an important factor since no participants mentioned any such thought process in the exit survey when describing their behavior.

negatively associated with their actual performance, and their evaluations of the risk assessment had no statistically significant relationship with its actual performance (other research has similarly shown that people struggle to detect algorithmic mistakes across a variety of conditions [400]). Given these results, it is no wonder that participants in the treatment group underperformed the risk assessment. How can we expect people to navigate the balance between their own judgment and a risk assessment's when they are unable to accurately assess their own or the algorithm's performance in the first place? Determining how to incorporate a risk assessment into one's own prediction is arguably a more challenging task that requires more expertise than merely making a prediction.

The results of this study raise one of the most important but rarely-discussed issues at the heart of debates about risk assessments: how *should* risk assessments be incorporated into existing practices? On the one hand, risk assessments alone achieve better performance than individuals (both with and without a risk assessment's aid) in terms of accuracy and false positive rates.[4] Yet there are many reasons to be wary of relying too heavily on risk assessments, including due process concerns, their embedding of discriminatory and punitive approaches to justice, and their potential to hinder more systemic criminal justice reforms [192, 460, 88] Meanwhile, the current approach of presenting predictions to judges without sufficient guidelines or training comes with the issues of poor interpretation and disparate interactions.

The conflicts between these positions are apparent in how the Wisconsin Supreme Court severely circumscribed the role of risk assessments in its decision in State v. Loomis, regarding the use of COMPAS in sentencing. Despite defending the use of COMPAS on the grounds that it "has the potential to provide sentencing courts with more complete information," the Court also mandated that "risk scores may not be used: (1) to determine whether an offender is incarcerated; or (2) to determine the severity of the sentence" [514]. If COMPAS is not supposed to influence the sentence, there are few purposes that the "more complete information" it provides

[4]This result assumes a comparison between a single individual and a risk assessment. This is in contrast to a recent study suggesting that humans are just as accurate as COMPAS: that result holds only when the predictions of humans are aggregated to create a "wisdom of the crowd" effect; in fact, that study similarly found COMPAS to be more accurate than individuals [141].

can serve—and few ways to ensure that it serves only those purposes. In that case, why show it at all?

With this in mind, the next chapter looks to the principles that are desirable when algorithms are incorporated into human decision-making processes.

Chapter 4

The Principles and Limits of Algorithm-in-the-Loop Decision-Making

4.1 Introduction

The emergence of novel algorithm-in-the-loop decision-making processes raises two questions—one normative, one empirical—that require answers before machine learning should be integrated into some of society's most consequential decisions:

1. What criteria characterize an ethical and responsible decision when a person is informed by an algorithm?
2. Do the ways that people make decisions when informed by an algorithm satisfy these criteria?

Both of these questions lack clear answers. While there exist many standards, policies, and studies related to the decisions made by people and institutions, our normative and empirical understanding of algorithm-in-the-loop decision-making is far thinner.

Despite widespread attention to incorporating ethical principles (most notably, fairness, accountability, and transparency) into algorithms, the principles required of the people using algorithms largely remain to be articulated and evaluated. For although calls to adopt machine learning models often focus on the accuracy of these tools [275, 99, 460, 340], accuracy is not only attribute of ethical and responsible decision-making. The principle of procedural justice, for instance, requires that decisions be (among other things) accurate, fair, consistent, correctable, and ethical [301]. Even as algorithms bear the potential to improve predictive accuracy, their inability to reason reflexively and adapt to novel or marginal circumstances makes them poorly suited to achieving many of these principles [11]. As a result, institutions implementing algorithmic advice may find themselves hailing the algorithm's potential to provide valuable information while simultaneously cautioning that the algorithm should not actually determine the decision that is made [514].

In practice, algorithm-in-the-loop decision-making requires synthesizing the often divergent capabilities of people and machine learning models. Despite this imperative, however, research and debates regarding algorithmic decision-making aids have primarily emphasized the models' statistical properties (e.g., accuracy and fairness) rather than their influence on human decisions [17, 130]. Thus, even as institutions increasingly adopt machine learning models in an attempt to be "evidence-based" [460, 291, 361, 498], relatively little is actually known about how machine learning models affect decision-making in practice. This lack of evidence is particularly troubling in light of research which suggests that people struggle to interpret machine learning models and to incorporate algorithmic predictions into their decisions, often leading machine learning systems to generate unexpected and unfair outcomes (see Chapter 3).

In this chapter, we explore both the normative and empirical dimensions of algorithm-in-the-loop decision-making. We focused on risk assessments—machine learning models that predict the probability of an adverse outcome—which are commonly used in algorithm-in-the-loop decisions in settings such as the criminal justice system.

We began by articulating a framework with which to evaluate human-algorithm interactions, positing three

desiderata that are essential to effective and responsible decision-making in algorithm-in-the-loop settings. These principles relate to the accuracy, reliability, and fairness of decisions. Although certainly not comprehensive, these desiderata provide a starting point on which to develop further standards for algorithm-in-the-loop decision-making.

We then ran experiments using Amazon Mechanical Turk to study whether people satisfy these principles when making predictions about risk. We explored these decisions in two settings where risk assessments are increasingly being deployed in practice—pretrial release hearings and financial loan applications [291, 361, 450]—and under several conditions for presenting the risk assessment or structuring the human-algorithm interaction. This experimental setup allowed us to evaluate algorithm-in-the-loop decision-making as a function of risk assessment presentation and to compare outcomes across distinct prediction tasks. Although these experiments involved laypeople rather than practitioners (such as judges or loan officers), meaning that we cannot take the observed behaviors to be a direct indication of how risk assessments are used in real-world settings, our results highlight potential challenges that must be factored into considerations of risk assessments.

People's behavior in the experiments reliably satisfied only one of our three principles for algorithm-in-the-loop decision-making. While almost every treatment improved the accuracy of predictions, no treatment satisfied our criteria for reliability and fairness. In particular, we found that under all conditions in both settings our study participants 1) were unable to effectively evaluate the accuracy of their own or the risk assessment's predictions, 2) did not calibrate their reliance on the risk assessment based on the risk assessment's performance, and 3) exhibited racial bias in their interactions with the risk assessment. Further research is necessary to determine whether the practitioners who use risk assessments exhibit similar behaviors.

4.2 Principles for Algorithm-in-the-Loop Decision-Making

An algorithm-in-the-loop framework provides a new approach to studying algorithmic decision-making aids: rather than evaluating models like risk assessments simply as statistical tools of prediction, we must consider them as sociotechnical tools that take shape only as they are integrated into social contexts (see Chapter 2). In other words, risk assessments are technologies of "social practice" that "are constituted through and inseparable from the specifically situated practices of their use" [469]. This means that a risk assessment's statistical properties (e.g., AUC and fairness) do not fully determine the risk assessment's impacts when introduced in social contexts. Given that the outcomes are ultimately more important than the statistical properties, a greater emphasis on the relationship between risk assessments and their social impacts is necessary.

Although arguments in favor of risk assessments often focus on the predictive accuracy of these tools [275, 99, 460, 340], many important decisions require more than just accuracy. For example, the principle of procedural justice requires that decisions be (among other things) accurate, fair, consistent, correctable, and ethical [301]. While many institutions have a long history of pursuing these goals and creating procedures to ensure that they are satisfied, achieving these goals in algorithm-in-the-loop settings requires new definitions, designs, and evaluations. Notably, although algorithms often make more accurate predictions than people do, their inability to reason reflexively and adapt to novel or marginal circumstances makes them poorly suited to achieving many principles of responsible and ethical decision-making [11]. Algorithm-in-the-loop decision-making thus requires synthesizing the often divergent capabilities of people and machine learning models.

As a starting point toward this end, we suggest three principles of behavior that are desirable in the context of making predictions (or decisions based on predictions) with the aid of machine learning models. Our three desiderata are as follows:

Desideratum 1 (Accuracy). People using the algorithm should make more accurate predictions than they could without the algorithm.

Desideratum 2 (Reliability). People should accurately evaluate their own and the algorithm's performance and should calibrate their use of the algorithm to account for its accuracy and errors.

Desideratum 3 (Fairness). People should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

Desideratum 1 is the most straightforward: the goal of introducing algorithms is typically to improve predictive performance [275, 99, 460, 340].

Desideratum 2 is important for algorithm-in-the-loop decision-making to be reliable, accountable, and fair. If people are unable to determine the accuracy of their own or the algorithm's decisions, they will not be able to appropriately synthesize these predictions to make reliable decisions. Such evaluation is essential to correcting algorithmic errors: "overriding" the risk assessment is commonly recognized as an essential feature of responsible decision-making with risk assessments [514, 243, 291, 498]. This principle is also important to ensuring the fairness of decisions, since algorithms are prone to making errors on the margins [11] and minority groups are often less well represented in datasets. Moreover, if people are unable to evaluate their own or an algorithm's decisions, they may feel less responsible and be held less accountable for the decisions they make.

Finally, Desideratum 3 connects to fundamental notions of fairness: decisions should be made without prejudice related to attributes such as race and gender. This is particularly important to consider given evidence that people engage in disparate interactions when making decisions with the aid of a risk assessment (see Chapter 3).

These three principles guided our analyses of the experimental results: we evaluated the participant behaviors according to each desideratum, demonstrating how all three can be quantitatively evaluated.

4.3 Methods

4.3.1 Study Design

See Section 2.4 for the full details of the study design. Here, I describe the elements of this experiment that are particular to this study.

Data and Risk Assessments

For every defendant and applicant, we used the `xgboostExplainer` package to determine the log-odds influence of each attribute on the risk assessment's predictions [170]. We selected samples of 300 defendants and applicants whose profiles would be shown to participants during the Mechanical Turk experiments (Table 4.1 and Table 4.2).

For the loans setting, we used records about all 421,095 loans issued during 2015. This yielded a dataset of 206,913 issued loans (Table 4.2). The average loan was for $15,133.51; the average applicant had an income of $78,093.47 and a "Good" credit score. Approximately three-quarters of these loans were fully paid.

Experiment Setup

After being sorted into either the pretrial or loans setting, participants were then randomly sorted into one of six conditions:

Baseline. Participants were presented with the narrative profile, without any information regarding the risk assessment. This condition represents the status quo prior to risk assessments, in which people made decisions without the aid of algorithms, and was one of our two control conditions.

RA Prediction. Participants were presented with the narrative profile as well as the risk assessment's prediction in simple numeric form. This condition represents the simplest presentation of a risk assessment and the typical risk assessment status quo, in which the advice of a model is presented in numerical or categorical

	Sample N=300	Black N=178	White N=122
Background			
Male	85.7%	87.6%	82.8%
Black	59.3%	100.0%	0.0%
Mean age	27.7	27.4	28.2
Drug crime	44.3%	49.4%	36.9%
Property crime	36.0%	32.0%	41.8%
Violent crime	14.7%	14.0%	15.6%
Public order crime	5.0%	4.5%	5.7%
Prior arrest(s)	55.0%	66.9%	37.7%
# of prior arrests	3.6	4.6	2.2
Prior conviction(s)	39.7%	50.0%	24.6%
# of prior convictions	2.2	2.8	1.3
Prior failure to appear	23.7%	30.3%	13.9%
Outcomes			
Rearrest	19.0%	24.2%	11.5%
Failure to appear	23.3%	28.1%	16.4%
Violation	32.3%	39.9%	21.3%

Table 4.1: Summary statistics for the 300 defendants presented to participants. See Table 2.1 for the full data sample.

	All N = 206,913	Sample N = 300
Applicant		
Annual income	$78,093.47 ($73,474.56)	$83,190.08 ($83,681.52)
Credit score	695.3 (30.5)	693.9 (30.3)
Home owner	10.2%	10.0%
Renter	40.1%	40.3%
Has mortgage	49.7%	49.7%
Loan		
Loan amount	$15,133.51 ($8,575.05)	$15,377.75 ($8,520.84)
36 months to pay off loan	70.5%	73.3%
Monthly payment	$448.49 ($251.44)	$462.19 ($253.86)
Interest rate	12.9% (4.5%)	13.05% (4.5%)
Outcomes		
Charged off	25.9%	26.0%

Table 4.2: Summary statistics for all approved loans in 2015 and for the 300-loan sample used in the Mechanical Turk experiments. Numbers in parentheses represent standard deviations.

form as a factor for the human decision maker to consider. This treatment served as the second control condition against which we evaluated the following four treatments, which represent a core (though not exhaustive) set of potential reforms to algorithmic decision aids.

Default. Participants were presented with the RA Prediction condition, except that the prediction form was automatically set to the risk assessment's prediction (Figure 4.1). Participants could select any desired value, however. In Chapter 3, many people followed this strategy when making predictions with the aid of a risk assessment, looking at the algorithm's prediction first and then considering whether to deviate from that value. Moreover, this condition accords with the implementations of risk assessments that treat the model's prediction as the presumptive default and require judges to justify any overrides [92, 498].

Update. Participants were first presented with the Baseline condition; after making a prediction, participants were presented with the RA Prediction condition (for the same case) and asked to make the prediction again. In Chapter 3, many people first made a prediction by themselves and then took the algorithm into model when making decisions with the aid of a risk assessment. This treatment adds structure to the prediction process (by prompting people to focus on the narrative profile before considering the risk assessment's prediction), which prior research has found improves decision-making [259, 300].

Explanation. Participants were presented with the RA Prediction condition along with an explanation that indicated which features made the risk assessment predict notably higher or lower levels of risk (Figure 4.1).[1] This treatment follows from the many calls to present explanations of machine learning predictions [137, 417, 148]. In addition, by indicating which attributes strongly influenced the risk assessment's prediction, this treatment may prevent people from double counting features that the model had already considered, a problem found in Chapter 3.

Feedback. Participants were presented with the RA Prediction condition; after submitting each prediction,

[1]The explanations were derived from the log-odds influence of each factor, with a threshold of 0.1 and -0.1 to be included in the lists of positive and negative factors, respectively.

participants were presented with an alert indicating the outcome of that case (e.g., whether the loan applicant actually defaulted on their loan). Although in practice immediate feedback on the outcomes of pretrial release or financial loans would not be available, this treatment provides one form of training for the users of machine learning systems, which is often regarded as an essential ingredient for the effective implementation of risk assessments [73, 243, 498].

In all settings and conditions, participants were presented with narrative profiles about a sample of 40 people drawn from the 300-person sample populations and were asked to predict their outcomes. Figure 4.1 presents examples of the prompts presented to participants when making predictions.

4.3.2 Analysis

We analyzed the behavior of participants using metrics related to three topics: the quality of participant predictions, the influence of the risk assessment on participant predictions, and the extent to which participants exhibited bias when making predictions.

Prediction performance measures

The first set of metrics evaluated the quality of participant predictions across treatments.

We evaluated the quality of each prediction using the Brier score. When presented with a loan applicant who does not default on their loan, for example, a prediction of 0% risk would yield a score of 1, a prediction of 100% would yield a reward of 0, and a prediction of 50% would yield a score of 0.75.

We defined the "participant prediction score" as the average Brier score attained among the 40 predictions that each participant made. Similarly, the "risk assessment prediction score" is the average Brier score attained by the risk assessment. These two metrics were used to evaluate the performance of each participant and the risk assessment.

We defined the performance gain produced by each treatment t as the improvement in the participant pre-

diction score achieved by participants in treatment t over participants in the Baseline condition, relative to the performance of the risk assessment:

$$Gain_t = \frac{S_t - S_B}{S_R - S_B} \tag{4.1}$$

where S_t, S_B, and S_R represent the average prediction scores of participants in the treatment t, of participants in Baseline, and of the risk assessment, respectively. By definition, the gain of the Baseline condition is 0 and the gain of the risk assessment is 1.

Risk assessment influence measures

The second set of metrics evaluated how much the risk assessment influenced participant predictions.

We measured the influence of the risk assessment by comparing the predictions made by participants who were shown the risk assessment with the predictions about the same case made by participants who were not shown the risk assessment. That is, the influence of the risk assessment on the prediction p_i^k by participant k about case $i \in \{1, \ldots, 300\}$ is

$$I_i^k = \frac{p_i^k - b_i}{r_i - b_i} \tag{4.2}$$

where b_i is the average prediction about that case made by participants in the Baseline treatment and r_i is the prediction about that case made by the risk assessment. For participants in Update, b_i is b_i^k: participant k's initial prediction about case i before being shown the risk assessment's prediction. This is akin to the "weight of advice" metric that has been used in other contexts to measure how much people alter their decisions when presented with advice [519, 309]. To obtain reliable measurements, when evaluating risk assessment influence we excluded all predictions for which $|r_i - b_i| < 0.05$.

Given an influence I_i^k, we can express each prediction as a weighted sum of the risk assessment and baseline predictions, where $p_i^k = (1 - I_i^k)b_i + I_i^k r_i$. $I = 0$ means that the participant ignored the risk assessment, $I = 0.5$ means that the participant equally weighed their initial prediction and the risk assessment, and $I = 1$ means that

the participant relied solely on the risk assessment.

Disparate interaction measures

The third set of metrics evaluated whether participants responded to the risk assessment in a racially biased manner. Following Chapter 3, we evaluated "disparate interactions" by comparing the behaviors of participants when making predictions about Black and white criminal defendants.[2] We measured disparate interactions in two ways.

Our first measure of disparate interactions compared the influence of the risk assessment on predictions made about Black and white defendants. We divided the data based on whether the risk assessment prediction r_i was greater or less than the baseline prediction b_i (and thus whether the risk assessment was likely to pull participants toward higher or lower predictions of risk). For each of these two scenarios, we measured the risk assessment's influence on predictions about Black defendants and white defendants; for example, we defined the influence on predictions about Black defendants when $r_i > b_i$ as $I_{Black,>} = \text{mean}\{I_i^k | \forall k, Race_i = \text{Black}, r_i > b_i\}$. We then defined the *RA influence disparity* as follows:

$$RA\ influence\ disparity_{>} = I_{Black,>} - I_{white,>} \tag{4.3}$$

$RA\ influence\ disparity_{>} > 0$ means that when $r_i > b_i$, participants were more strongly influenced to increase their predictions of risk when evaluating Black defendants than when evaluating white defendants.

Our second measure of disparate interactions compared the extent to which participants deviated from the risk assessment's suggestion when making predictions. For each prediction p_i^k by participant k about defendant i, we measured the participant's deviation from the risk assessment as $d_i^k = p_i^k - r_i$ (i.e., $d_i^k > 0$ means that participant k predicted a higher level of risk than the risk assessment about defendant i). We used this metric to

[2]Because we did not possess demographic characteristics about the loan applicants, we applied this analysis only to the pretrial setting.

measure the average deviation for each race; for example, the average deviation for all predictions about Black defendants is $D_{Black} = \text{mean}\{d_i^k | \forall k, Race_i = \text{Black}\}$. We then defined the *Deviation disparity* as follows:

$$Deviation\ disparity = D_{Black} - D_{white} \tag{4.4}$$

Deviation disparity > 0 means that participants were more likely to deviate positively when evaluating Black defendants than when evaluating white defendants.

4.4 Results

We conducted trials on Mechanical Turk over the course of several weeks in March 2019. Filtering out workers who failed at least one of the attention check questions, who required more than three attempts to pass the comprehension test, and who participated in the experiment more than once[3] yielded a population of 1156 participants in the pretrial setting and 732 participants in the loans setting (Table **??**). Across both settings, a majority of participants were male, white, and have completed at least a college degree. We asked participants to self-report their familiarity with the U.S. criminal justice system, financial lending, and machine learning on a Likert scale from "Not at all" (1) to "Extremely" (5). The average reported familiarity with the three topics in each setting was between "Slightly" (2) and "Moderately" (3), with little variation across treatments.

Participants reported in the exit survey that the experiment paid well, was clear, and was enjoyable. Considering both the base payment and the bonus payment, participants in the pretrial setting earned an average wage of $15.20 per hour and participants in the loans setting earned an average wage of $17.18 per hour. Out of 213 participants who responded to a free text question in the exit survey asking for any further comments, 32% mentioned that the experiment length and payment were fair. Participants were also asked in the exit survey to rate how clear and enjoyable the experiment was on a Likert scale from "Not at all" (1) to "Extremely" (5). More

[3] A server load issue prevented us from recognizing all repeat users when they entered the experiment.

than 90% of participants in both settings reported that the experiment was "Very" or "Extremely" clear, and more than half of participants in both settings stated that the experiment was "Very" or "Extremely" enjoyable.

	Pretrial N=1156	**Loans** N=732
Demographics		
Male	55.3%	53.0%
Black	7.1%	7.2%
White	77.2%	77.6%
18-24 years old	8.4%	7.9%
25-34 years old	42.4%	44.5%
35-59 years old	45.0%	43.2%
60+ years old	4.2%	4.4%
College degree or higher	70.9%	71.7%
Criminal justice familiarity	2.8	2.9
Financial lending familiarity	2.7	2.9
Machine learning familiarity	2.4	2.5
Treatment		
Baseline	16.5% (N=191)	15.3% (N=112)
Risk Assessment	17.3% (N=200)	16.9% (N=124)
Default	16.9% (N=195)	17.6% (N=129)
Update	16.1% (N=186)	17.9% (N=131)
Explanation	15.1% (N=174)	16.8% (N=123)
Feedback	18.2% (N=210)	15.4% (N=113)
Outcomes		
Participant hourly wage	$15.20	$17.18
Experiment clarity	4.4	4.4
Experiment enjoyment	3.5	3.7

Table 4.3: Attributes of the participants in our experiments.

In response to exit survey questions asking how they made predictions, participants reported a variety of strategies for using the risk assessment:

- Follow the risk assessment in most or all cases (e.g., "i mostly trusted the algorithm to be more objective than i was.").
- Use the risk assessment as a starting point and then adjust based on the narrative profile (e.g., "It served

as a jumping off point for my prediction.").

- Rely on the risk assessment only when unsure about a particular prediction (e.g., "I put my trust into the algorithm's predictions for when I felt like I wasn't too sure.").
- Make a prediction without the risk assessment and then adjust based on the risk assessment (e.g., "I tried not to look at it until I came to my own conclusion and then I rated my score against the computers.").
- Ignore the risk assessment (e.g., "I don't think the algorithm can be relied on").

Participants in the pretrial setting also reported diverging approaches with regard to race: while 4.4% of participants reported that they considered race when making predictions, 2.2% of participants reported explicitly ignoring race. These opposing strategies reflect differences in the perceived relationship between race and prediction: participants in the first category saw race as a factor that could improve their predictive accuracy, while participants in the second category saw race as a factor that should not be incorporated into predictions of risk (e.g., "I tried to ignore race").

4.4.1 Desideratum 1 (Accuracy)

Desideratum 1 states that people using the algorithm should make more accurate predictions than they could if working alone. We found that every treatment except Feedback reliably improved performance over the Baseline treatment and that the Update treatment yielded the best performance across both settings.

Across all predictions in the pretrial setting, the average participant prediction score was 0.768 and the average risk assessment prediction score was 0.803. Aside from Feedback (whose performance was not statistically distinct from that of Baseline), every treatment yielded a performance that was statistically significantly greater than Baseline and lower than the risk assessment. Compared to RA Prediction, which had an average prediction score of 0.774, two treatments (aside from Baseline) had statistically significant differences: Feedback had a lower average prediction score of 0.751 ($p < 10^{-6}$, Cohen's $d = 0.08$), while Update had a higher average score of

0.782 ($p = 0.041$, $d = 0.03$). The gain produced by each non-Baseline treatment (Equation 4.1) ranged from 0.011 for Feedback to 0.603 for Update, while RA Prediction achieved a gain of 0.464 (Figure 4.2). Update produced a prediction score that was 1.0% greater and a gain that was 30.0% larger than RA Prediction.

A similar pattern emerged in the loans setting. Across all predictions in the pretrial setting, the average participant prediction score was 0.793 and the average risk assessment prediction score was 0.823. Compared to RA Prediction, which had an average prediction score of 0.802, two treatments (aside from Baseline) had statistically significant differences: Feedback had a lower average prediction score of 0.779 ($p < 10^{-4}$, $d = 0.09$), while Update had a higher average score of 0.813 ($p = 0.019$, $d = 0.05$). The gain produced by each non-Baseline treatment ranged from 0.327 for Feedback to 0.821 for Update, while RA Prediction achieved a gain of 0.682 (Figure 4.2). In other words, Update produced a prediction score that was 1.4% greater and a gain that was 20.4% larger than RA Prediction.

The relative performance of each treatment was similar across the two settings (Figure 4.2): the gain of the five non-Baseline treatments had a Pearson correlation of 0.96 ($p = 0.010$) and a Spearman correlation of 0.9 ($p = 0.083$). In both settings, Feedback yielded significantly worse performance than RA Prediction, while Update produced significantly better performance.

To evaluate the relationship between model performance and model presentation, we measured how much more or less accurate the risk assessment would have needed to be for RA Prediction to yield the same performance as the other treatments. Taking all of the predictions made by participants in RA Prediction, we regressed the participant prediction score on the risk assessment's prediction score to determine how participant performance depends on model performance. In both cases the slope was close to 1 (1.14 in pretrial, 0.98 in loans) and was significant with $p < 10^{-15}$. In the pretrial setting, Update was equivalent to RA Prediction with a risk assessment that performs 0.91% better than the actual risk assessment while Feedback was equivalent to RA Prediction with a risk assessment that performs 2.52% worse (a range of 3.43%). In the loans setting, Update was equivalent to RA Prediction with a risk assessment that performs 1.35% better than the actual risk assessment

while Feedback was equivalent to RA Prediction with a risk assessment that performs 2.91% worse (a range of 4.26%).

We observed several patterns that can partially account for the different performance levels observed. The average participant prediction score in each treatment was closely related to the rate at which participants matched their prediction to the risk assessment's prediction: the more often participants in a treatment followed the risk assessment's advice, the better the average participant prediction score in that treatment ($p = 0.012$ in pretrial, $p = 0.055$ in loans).

Although we were unable to ascertain clear explanations for why participants matched the risk assessment at different rates in every treatment, a striking pattern emerged in the Feedback treatment, which had by far the lowest match rate in both settings: the match rate declined drastically after the first prediction. In the pretrial setting, for example, the match rate of the first prediction in Feedback was 42.9%, whereas the match rate for the following 39 predictions ranged between 22.9% and 31.4% (average=26.4%). This was due to a shift in participant predictions toward the extremes (0% and 100%). For instance, the rate at which participants predicted 0% risk increased by a factor of 1.8 and 2.8 after the first prediction in the pretrial and loans settings, respectively. This indicates that many participants responded to the feedback presented after the first prediction (this feedback was necessarily binary, since the outcome either did or did not occur) by treating their own predictions as binary. This change in behavior led to a decrease in the performance of participants in the Feedback treatment.

We further analyzed the Update treatment by evaluating the quality of participants' initial predictions, which they made before being shown the risk assessment for that case. Surprisingly, despite making predictions under the same condition as participants in Baseline, participants' initial predictions in Update outperformed the predictions made in Baseline (pretrial: 0.772 vs. 0.750, $p < 10^{-5}$; loans: 0.799 vs. 0.757, $p < 10^{-14}$). This appeared to be due to the risk assessment serving a training role for participants: the initial predictions in Update improved over the course of the 40 predictions in the pretrial setting[4] ($p = 0.015$) and exhibited a sharp improvement after

[4]In only one other treatment across the two settings did participant performance improve statistically significantly over time.

the first prediction in the loans setting, suggesting that being shown an algorithm's prediction about some cases can help people make more accurate predictions about future cases. The final predictions in Update, made with the benefit of the risk assessment's advice, provided further improvement over the initial predictions (pretrial: 0.782 vs. 0.772, $p = 0.014$; loans: 0.813 vs. 0.799, $p = 0.002$). These results suggest that the improvement produced by the Update treatment was twofold: first, it trained participants to make more accurate predictions in general, and second, it provided the risk assessment's prediction for the particular case at hand.

4.4.2 Desideratum 2 (Reliability)

Desideratum 2 states that people should accurately evaluate their own and the algorithm's performance and should calibrate their use of the algorithm to account for its accuracy and errors. This principle involves two components: first, the ability to evaluate performance, and second, the ability to calibrate a decision based on the algorithm's performance. We found that participants could not reliably exhibit either of these behaviors in any treatment.

Evaluation

We assessed whether participants could evaluate their own and the risk assessment's performance by comparing participant exit survey responses to the actual behaviors that they exhibited and observed (Table **??**). Participants were asked to respond to each question on a Likert scale from "Not at all" (1) to "Extremely" (5).

To measure perceptions of their own performance, all participants were asked "How confident were you in your decisions?" We evaluated whether participants' self-reported confidence in their performance was related to their actual performance. The average participant confidence was 3.1 in pretrial and 3.2 in loans. Within each treatment in both settings, we regressed confidence on performance, controlling for each participant's demographic information and exit survey responses, along with the risk assessment's performance (Table **??**). Across both settings, the only statistically signifiant relationships between a participant's confidence and performance

emerged as negative negative associations in Default and Update in loans ($p = 0.03$ and $p = 0.047$, respectively). In none of the treatments could participants reliably evaluate their performance, in some cases actually performing less well as they became more confident.

To measure participant evaluations of the risk assessment's performance, we asked every participant who was shown the risk assessment "How accurate do you think the risk score algorithm is?" and analyzed whether participant responses reflected the risk assessment's accuracy.[5] The average report of algorithm accuracy was 3.1 in pretrial and 3.3 in loans. Within each treatment in both settings, we regressed the participant evaluations of the risk assessment's accuracy against the risk assessment's actual performance, controlling for each participant's performance, demographic information, and exit survey responses (Table **??**). In the Update treatment in both settings ($p = 0.04$ in pretrial and $p < 10^{-3}$ in loans) and in the Default treatment in loans ($p = 0.01$), participant evaluations of the risk assessment were negatively associated with the risk assessment's actual performance. In no treatment or setting were participants able to accurately evaluate the risk assessment's performance.

Calibration

To evaluate whether participants calibrated their use of the risk assessment to the risk assessment's performance, we compared the influence of the risk assessment on each prediction (Equation 4.2) with the quality of the risk assessment's predictions. Within each treatment, we regressed the risk assessment's influence on each participant prediction on the risk assessment's score for that prediction (Table **??**). Across all settings and treatments, only the Explanation treatment in the loans setting had a positive and statistically significant relationship in which people relied more strongly on the risk assessment as its performance improved ($p = 0.006$); in pretrial, however, Explanation, RA Prediction, and Feedback had a negative relationship in which people relied less strongly on the risk assessment as its performance improved ($p \leq 0.04$). In the six other treatments across the two settings,

[5] Although all participants were presented with predictions from the same model, each participant was presented with a different set of 40 predictions. As a result of this variation, each participant observed a different level of risk assessment quality.

	Confidence		RA Accuracy		Calibration	
	Pretrial	Loans	Pretrial	Loans	Pretrial	Loans
RA Prediction	0	0	0	0	-	0
Default	0	-	0	-	0	0
Update	0	-	-	-	0	0
Explanation	0	0	0	0	-	+
Feedback	0	0	0	0	-	0

Table 4.4: Summary of participant abilities to evaluate performance (first two columns) and to calibrate their predictions (third column). The columns measure the relationships between between participant confidence and actual performance (Confidence), participant estimates of the algorithm's performance and its actual performance (RA Accuracy), and participant reliance on the risk assessment and the risk assessment's performance (Calibration). + signifies a positive and statistically significant relationship, - signifies a negative and statistically significant relationship, and 0 signifies no statistically significant relationship. In all cases, + means that the desired behavior was observed.

participants did not differentiate their reliance on the risk assessment based on how it actually performed.

4.4.3 Desideratum 3 (Fairness)

Desideratum 3 states that people should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes.

To assess whether this desideratum was satisfied, we analyzed if any "disparate interactions" emerged in the various treatments. Because Desideratum 3 concerns bias with respect to sensitive attributes and the loans data did not contain any such attributes about applicants, we applied this analysis only in the pretrial setting. We analyzed disparate interactions along two framings: first, comparing the risk assessment's influence on participants when making predictions about Black and white defendants, and second, comparing the participant deviations from the risk assessment when making predictions about Black and white defendants. In both cases, we found that every treatment exhibited disparate interactions and that the Update treatment yielded the smallest disparate interactions.

Influence of the risk assessment

For each treatment, we compared the influence of the risk assessment on predictions about Black and white defendants (Equation 4.3). We broke down the analysis based on whether the risk assessment's prediction was greater or less than the average Baseline participant prediction for that defendant ($r_i > b_i$ and $r_i < b_i$, respectively).

In cases where $r_i > b_i$, the risk assessment exerted a larger influence to increase risk on predictions about Black than white defendants in every treatment (Figure 4.3). These differences were statistically significant in three of the five treatments: RA Prediction ($p = 0.001$), Update ($p < 10^{-4}$), and Feedback ($p = 0.02$). The largest disparities of 0.38 occurred in Feedback and RA Prediction; in the latter, for example, the influence for Black defendants was 0.50 (meaning that participants equally weighed their own and the risk assessment's judgments) and the influence for white defendants was 0.12 (meaning that participants only slightly considered the risk assessment's judgments). The smallest disparity of 0.07 occurred in Update. Thus, although the *RA influence disparity*$_>$ was positive in Update, the disparity was reduced by 81.5% compared to RA Prediction.

The inverse pattern emerged in cases where $r_i < b_i$: in every treatment, the risk assessment exerted a greater influence to reduce risk when participants were evaluating white defendants. The discrepancies between Black and white defendants were reduced, however, and were significant only in the Update treatment, which had a disparity of 0.05 ($p = 0.02$).

Deviation from the risk assessment

For each treatment, we compared the extent to which participants deviated from the risk assessment when making predictions about Black versus white defendants (Equation 4.4). In every treatment, participants on average deviated positively (toward higher risk) for Black defendants and negatively (toward lower risk) for white defendants. Aside from Update ($p = 0.053$), these deviation disparities were statistically significant in every treatment ($p < 10^{-6}$). The largest gap in average deviations (of 4.1%) came in Feedback, where the average deviation was +1.3% for Black defendants and -2.8% for white defendants. The smallest disparity (of 0.6%)

came in Update, where the average deviation was +0.4% for Black defendants and -0.2% for white defendants. Compared to RA Prediction, which had a disparity of 2.3%, Update reduced the *Deviation disparity* by 73.9%.

4.5 Discussion

This study explored the normative and empirical dimensions of algorithm-in-the-loop decision-making, with a focus on risk assessments in pretrial adjudication and financial lending. We first posited three desiderata as essential to facilitating accurate, reliable, and fair algorithm-in-the-loop decision-making. We then ran experiments to evaluate whether people met the conditions of these principles when making decisions with the aid of a machine learning model. We studied how people made predictions in two distinct settings under six conditions—including four that follow proposed approaches for presenting risk assessments—and found that only the desideratum related to accuracy was satisfied by any treatment. No matter how the risk assessment was presented, participants could not determine their own or the model's accuracy, failed to calibrate their use of the model to the quality of its predictions, and exhibited disparate interactions when making predictions.

These results call into question foundational assumptions about the efficacy and reliability of algorithm-in-the-loop decision-making. It is often assumed that, because risk assessments are merely decision-making aids, the people who make the final decisions will provide an important check on a model's predictions [514, 243, 291]. For example, in *State v. Loomis*, the Wisconsin Supreme Court mandated that COMPAS should be accompanied by a notice about the model's limitations and emphasized that staff and courts should "exercise discretion when assessing a COMPAS risk score with respect to each individual defendant" [514]. But such behavior requires people to evaluate the quality of predictions and to calibrate their decisions based on these evaluations—abilities that our findings indicate people do not reliably possess. That assumptions about human oversight are so central to risk assessment advocacy and governance is particularly troubling given the inability of algorithms to reason about novel or marginal cases [11]: people may make more accurate predictions on average when informed by

an algorithm, but they are unlikely to recognize and discount any errors that arise. Even when people are making the final decisions, using a risk assessment may reduce the capacity for reflexivity and adaptation within the decision-making process. These concerns are particularly salient given the persistence of disparate interactions across all of our experimental treatments.

The first step toward remedying these issues is to further develop criteria that should govern algorithm-in-the-loop decision-making. If society is to trust the widespread integration of machine learning models into high-stakes decisions, it must be confident that the decision-making processes that emerge will be ethical and responsible. Rather than emphasizing only those values which technology is capable of promoting (such as accuracy), society must evaluate technology according to a full slate of normative and political considerations, paying particular attention to the technology's downstream implications [191, 194]. Despite providing initial steps in this direction, the three desiderata proposed here are not comprehensive and may not even be of primary concern in certain contexts. Our three desiderata do not capture broader considerations such as whether the context of a decision is just and whether it is appropriate to incorporate algorithmic advice into that context at all. Existing theories of justice must be more thoroughly adapted to algorithm-in-the-loop decision-making and to the contexts in which these decisions arise.

Another important step will be to develop a deeper science of human-algorithm interactions for decision-making. Although debates about risk assessments have centered on the statistical properties of the models themselves [17, 130], we found that varying risk assessment presentation and structure affected the accuracy of human decisions to an extent equivalent to altering the underlying risk assessment accuracy by more than 4%. The relative performance of each treatment was similar across two distinct domains, suggesting that our results may reflect general patterns of human-algorithm interactions. But while we were able to explain some of the differences in treatment performance, we lack a comprehensive understanding of how risk assessment presentation affected people's behaviors. Notably, we found several counterintuitive results that challenge assumptions about how to improve human-algorithm interactions. Although it is commonly assumed that providing explanations

will improve people's ability to understand and take advantage of an algorithm's advice [148, 137, 417], we found that explanations did not improve human performance, a result that accords with prior work [400, 357]. We also found, counterintuitively, that providing feedback to participants significantly decreased participant accuracy (in one setting leading to predictions that were no better than those made without the advice of a risk assessment at all) and exacerbated disparate interactions.

More broadly, evaluations of algorithm-in-the-loop decision-making should consider not just the quality of decisions (the focus of this study) but also how working with an algorithm can change one's perceptions of the task itself. The presentation of models can shape people's responses to the predictions made, prompting people to focus on the predictive dimensions of a complex decision and suggesting particular assumptions. For example, predictive policing systems have prompted police to alter their focus while on patrol [45, 238] and are sometimes displayed in a manner that could exacerbate a militaristic police mindset [194].

The presentation and structure of an algorithm could also diminish someone's sense of moral agency when making predictions. Prior work has found that using automated systems can generate a "moral buffer" that prompts people to feel less responsible and accountable for their actions [112]. For behavior within algorithm-in-the-loop settings to be reliable and accountable, it is essential that human decision makers feel responsibility for their actions rather than deferring agency to the computer. As a corollary, in the face of "moral crumple zones" that place undue responsibility on the human operators of computer systems rather than on the creators of those systems [149], the people developing algorithmic decision aids must feel responsibility and be accountable for how their design choices affect the final decision makers' actions.

With these considerations in mind, an important direction of future work will be to develop design principles for algorithms—as well as for the social and political contexts in which they are embedded—to promote reliable, fair, and accountable decision-making. Given that only the accuracy desideratum was satisfied even when various interventions were tested, a great deal of work is clearly required to promote the full slate of desired behaviors. Such work requires a fundamental shift in algorithmic practice that begins with expanding the goals of develop-

ment and evaluation to include considerations beyond model accuracy. Producing algorithms for use in social contexts means not just designing technology, but designing sociotechnical systems in which human-algorithm interactions, governance, and political discourse are all as central to the outcomes as the model predictions themselves. A thorough understanding of how each of these factors affects the impacts of algorithms is essential to building sociotechnical systems that can reliably produce ethical outcomes.

This study was hindered by the limits of its methodology and scope. Our experiments abstracted human decision-making into a series of prediction tasks, thus potentially overstating the importance of accuracy and removing many other important factors from consideration. In the U.S. criminal justice system, for instance, decisions must satisfy due process and equal protection, meaning that defendants must have the right to hear and challenge claims against them, that rules based on accurate statistical generalizations are often rejected in favor of treating people like individuals, and that decisions must be made without discriminatory intent. Because these considerations were not captured by our experimental task or evaluation metrics, experiments such as ours—by nature of how they are designed—fail to provide a holistic evaluation of risk assessments' merits and flaws. Thus, even as future work further develops principles and methods for ethical algorithm-in-the-loop decision-making, it is necessary to retain a focus on the broader questions of justice that surround human-algorithm interactions and algorithmic policy interventions. The next chapter provides a first step toward this broader approach, extending this experimental algorithm-in-the-loop methodology to explore human *decisions* rather than only human *predictions*.

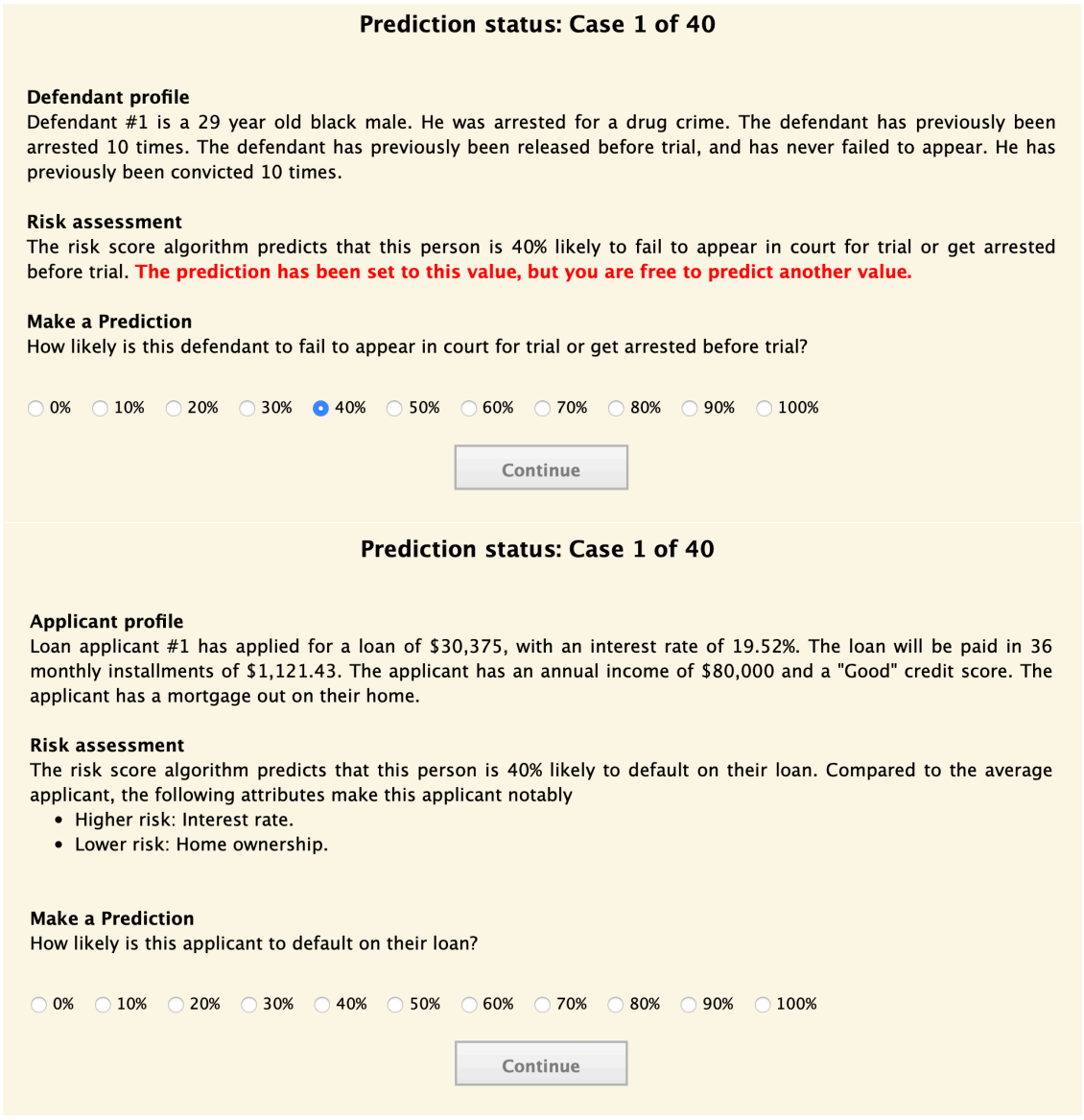

Figure 4.1: Examples of the prompts presented to participants in two of the six treatments. The top example is from the Default treatment (note that the "40%" bubble is already filled in, following the risk assessment's prediction) in the pretrial setting, while the bottom example is from the Explanation treatment in the loans setting.

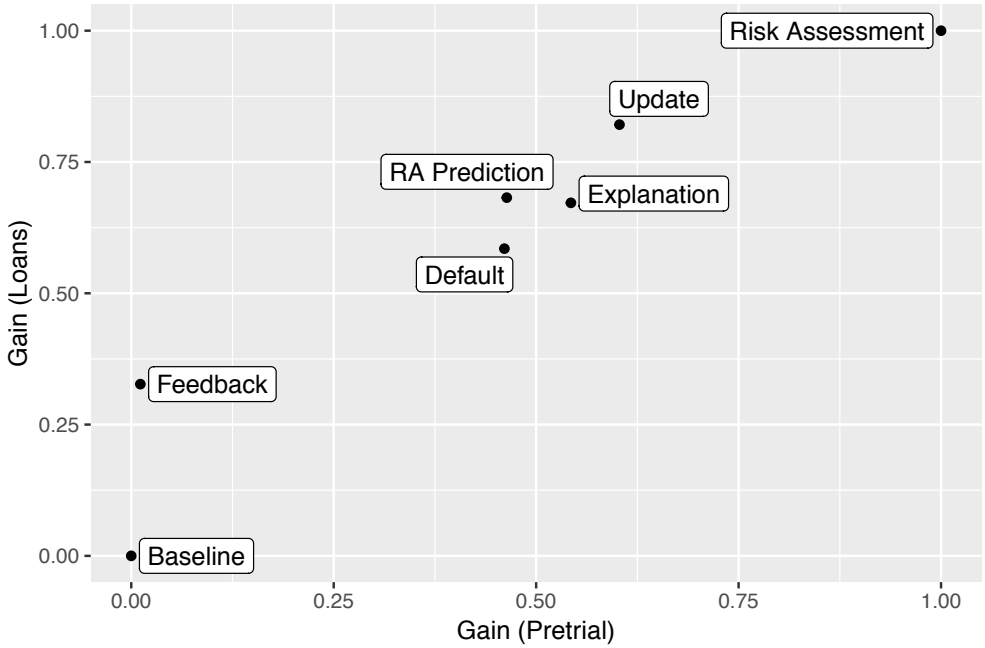

Figure 4.2: The relative performance gain (Equation 4.1) achieved by each experimental condition across the pretrial and loans settings. In both settings, the Update treatment performed statistically significantly better than RA Prediction and the Feedback treatment performed statistically significantly worse. Across the two settings, the gain of the conditions was highly correlated.

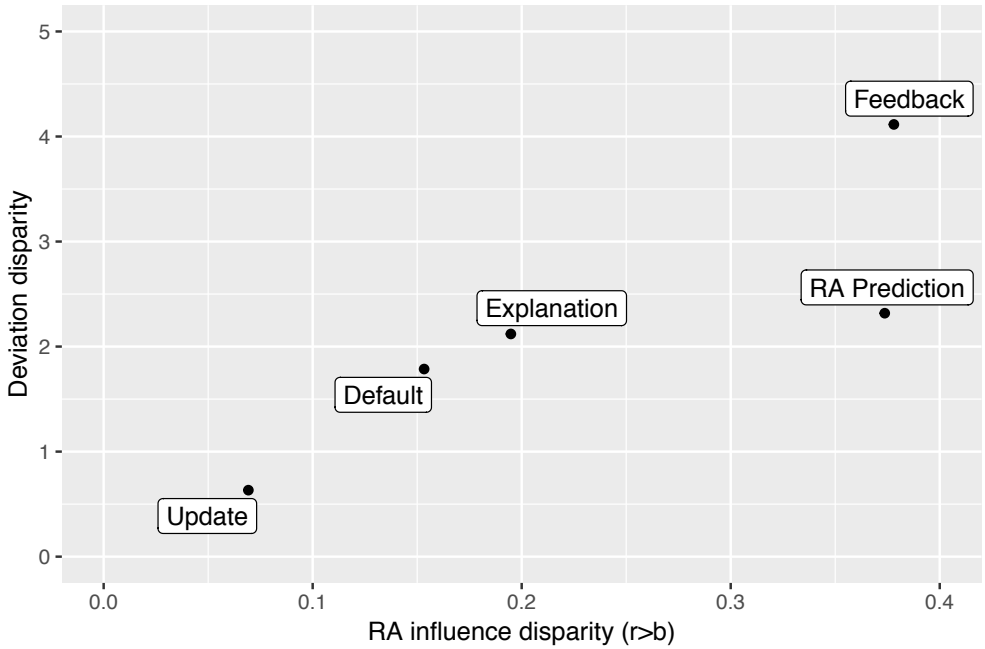

Figure 4.3: The disparate interactions present in each treatment in the pretrial setting, measured by the disparities in risk assessment influence (Equation 4.3) and in participant deviations (Equation 4.4) for Black versus white defendants. In both cases, values closer to 0 indicate lower levels of bias. The Update treatment yielded the smallest disparate interactions along both metrics, reducing the disparities (compared to RA Prediction) by 81.5% and 73.9%, respectively.

Chapter 5

Algorithmic Risk Assessments Can Distort Human Decision-Making in High-Stakes Government Contexts

5.1 Introduction

Although claims about the benefits of public sector risk assessments tend to directly compare humans and algorithms [275] and to emphasize the benefits of algorithms improving predictions [276], in practice algorithmic predictions are used by people to make complex decisions such as whether to release or detain criminal defendants before their trials. Determining whether risk assessments improve policy outcomes therefore requires understanding how these algorithms affect human decision-making.

We categorize the influence of risk assessments into four "settings" based on their effects on predictions and decisions, as summarized in Table 5.1. Setting 1 represents the baseline condition without any risk assessment

	Decision-making process unaffected by RA	Decision-making process affected by RA
Prediction-making process unaffected by RA	Setting 1 (RA does not affect prediction-making or decision-making processes)	Setting 2 (RA affects only decision-making process)
Prediction-making process affected by RA	Setting 3 (RA affects only prediction-making process)	Setting 4 (RA affects both prediction-making and decision-making processes)

Table 5.1: The four possible "settings" of how prediction-making and decision-making can be affected by a risk assessment (RA). Setting 1 represents a baseline process in which both prediction-making and decision-making are unaffected by a risk assessment. Settings 2-4 represent the possible conditions when decision-makers are presented with and affected by a risk assessment. Setting 3 represents the scenario in which risk assessments influence prediction-making but not decision-making; while decisions may differ in Setting 3 compared to Setting 1, this would be due to shifts in predictions rather than to shifts in how people make decisions as a function of predictions. Given extensive evidence that risk assessments affect human predictions, Setting 2 is relatively implausible.

and Settings 2-4 represent the possible conditions that could result when decision-makers are presented with and influenced by a risk assessment.

In response to concerns about government use of algorithms to make consequential decisions, lawmakers and other officials typically state that risk assessments merely provide accurate predictions to aid human decision-makers, who retain autonomy and discretion to make final decisions [154, 361, 514, 460]. This response assumes that risk assessments improve predictions of risk and thus ground decisions in better information, but do not alter the decision-making process itself (this would represent a shift from Setting 1 to Setting 3). Yet recent research indicates that risk assessments may also alter decision-making in unintended and often harmful ways (this would represent a shift to Setting 4). For instance, contrary to expectations, the use of pretrial risk assessments has exacerbated racial disparities in pretrial detention, in part because judges make more punitive decisions in response to risk predictions when evaluating Black defendants [8, 107, 464, 466]. Experimental evidence has also demonstrated that risk assessments prompt judges and law students to prioritize reducing risk relative to other considerations when making sentencing decisions [451, 460]. Such evidence suggests that risk assessments may unexpectedly alter how fundamental conceptions of justice are applied in practice.

In this study we use an online experiment to test whether risk assessments merely provide accurate predictions to aid human decision-makers (Setting 3), as is commonly asserted, or also alter decision-making itself (Setting 4) in two high-stakes public sector settings: 1) a pretrial setting where decisions about whether to release or detain criminal defendants before their trial depend in part on the risk that defendants would fail to appear in court for trial or would be arrested before trial, and 2) a loans setting where decisions about whether to approve or reject applications for government home improvement loans depend in part on the risk that applicants would default on the loan.

Based on evidence that framing decisions around losses more strongly motivates decision-makers (including judges) to avoid those losses [410, 442, 485, 259], we hypothesize that people presented with the predictions of risk assessments, which emphasize the risk of particular adverse outcomes, will place more emphasis on reducing risk when making decisions. Such a shift in decision-making would be notable for two primary reasons. First, unlike an improved capacity for predicting risk, an increased emphasis on reducing risk in government decision-making amounts to a shift in public policy, yet would occur here as a byproduct of adopting a technical tool rather than through a democratic policymaking process. Second, because "risk" is intertwined with legacies of racial discrimination in both the criminal justice (see Chapter 8) and loans [274] settings studied here, more heavily basing decisions on risk can exacerbate racial disparities in punishment and government aid.

5.2 Methods

5.2.1 Study Design

See Section 2.4 for the full details of the study design. Here, I describe the elements of this experiment that are particular to this study.

	Sample N=300	Black N=189	White N=111
Background			
Male	86.7%	88.4%	83.8%
Black	63.0%	100.0%	0.0%
Mean age at arrest	28.1	27.1	29.8
Drug crime	49.3%	50.8%	46.8%
Property crime	30.3%	28.0%	34.2%
Violent crime	14.0%	14.3%	13.5%
Public order crime	6.3%	6.9%	5.4%
Has prior arrests?	64.7%	73.5%	49.5%
Mean number of prior arrests	4.3	5.0	3.1
Has prior convictions?	50.0%	57.7%	36.9%
Mean number of prior convictions	2.4	2.9	1.7
Has prior failure to appear?	31.7%	34.4%	27.0%
Outcomes			
Rearrest	19.0%	20.1%	17.1%
Failure to appear	25.3%	29.6%	18.0%
Violation	36.0%	39.2%	30.6%

Table 5.2: Summary statistics for the 300 defendants presented to participants. See Table 2.1 for the attributes of the full data sample.

Data and Risk Assessments

The 300-person samples of defendants and applicants presented to participants in experiments are described in Table 5.2 and Table 5.3. In the loans setting, we restricted our analysis to loans that were issued specifically for home improvements between 2007 and 2018, which represents 6.7% of the total issued loans (the third most common purpose, following debt consolidation and paying off credit cards). This yielded a dataset of 45,218 issued home improvement loans. The average loan was for $14,556.38; the average applicant had an income of $95,262.88 and a credit score of 707.5 (categorized by FICO as "Good"). More than 80% of these loans were fully paid.

	All N=45,218	Sample N=300
Applicant		
Mean annual income	$95,262.88	$93,349.22
Mean credit score	707.5	705.9
Has "good" credit score?	65.7%	64.3%
Has mortgage?	83.9%	83.0%
Loan		
Mean loan amount	$14,556.38	$14,076.00
Mean months to pay off loan	42.4	42.6
Mean monthly payment	$435.75	$419.49
Mean interest rate	13.0%	13.2%
Outcome		
Loan paid off	83.2%	84.7%
Loan defaulted on	16.8%	15.3%

Table 5.3: Attributes of full sample of home improvement loans that were approved and the 300-loan sample presented to participants in experiments.

Experiment Setup

The experiment followed a 2x2x2 design, with splits along the following three dimensions (Figure 5.1):

- Pretrial release setting (50% of participants) vs. government home improvement loans setting (50%).
- Not presented with risk assessment (50%) vs. presented with risk assessment (50%). This is our primary experimental treatment.
- Decisions (75%) vs. predictions (25%).

Participants in the pretrial setting were required to make decisions or predictions about criminal defendants who have been arrested and are awaiting trial (Figure 5.2). Participants making decisions were tasked with choosing whether to detain or release 30 criminal defendants before their trials; participants making predictions were tasked with estimating the likelihood that 40 criminal defendants would (if released) be rearrested before trial or fail to appear in court for trial. Participants in the loans setting were required to make decisions or predictions

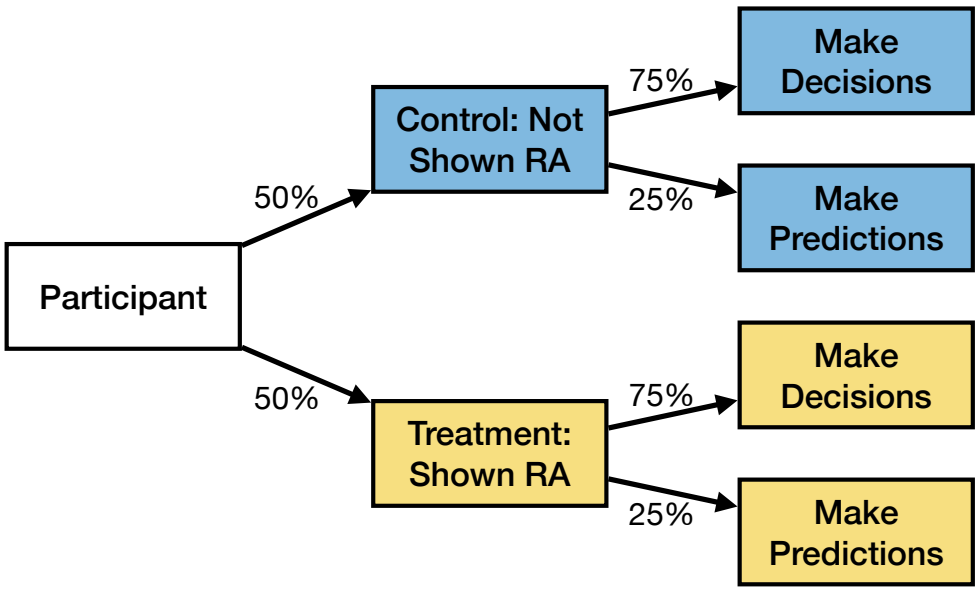

Figure 5.1: The four conditions that participants were sorted into in each setting, with probabilities indicating the likelihoods at each split. In each setting, every participant was sorted into one of the four terminal node scenarios. The first split is our primary experimental treatment: whether or not people are presented with the risk assessment. The second split enables us to estimate the perceived risk estimates of decision-makers without confounding the experiment by directly asking them to make predictions. In order to account for the effect of the risk assessment on predictions, the perceived risk measured for decisions in the control group are based only on predictions made in the control group and the perceived risk measured for decisions in the treatment group are based only on predictions made in the treatment group. Participants in all four scenarios were presented with the same set of 300 defendants or applicants.

about people who have applied for home improvement loans. Participants making decisions were tasked with choosing whether to approve or reject 30 loans; participants making predictions with tasked with estimating the likelihood that 40 loan applicants would (if granted a loan) default on their loans. In both settings, participants were presented with the narrative profiles of subjects drawn from the 300-person sample populations.

Because participant decisions are informed by their estimates of each defendant's or applicant's risk, we needed a measure of participants' estimates of risk about each subject. We could not directly ask participants making decisions for their estimates of risk, as doing so could prime them to consider risk, confounding the effect of showing the risk assessment. We therefore further split participants such that 25% were asked to make predictions of risk about 40 subjects drawn from the pools of 300 defendants or applicants (Figure 5.1).

We used these prediction-making participants to estimate the "perceived risk" of the decision-making par-

A Pretrial

Defendant Profile

The defendant is a 26 year old black male. He was arrested for a property crime. The defendant has previously been arrested 10 times. The defendant has previously been released before trial, and has never failed to appear. He has previously been convicted 10 times.

Risk Assessment Algorithm

The risk assessment algorithm predicts that this person is 40% likely to fail to appear in court for trial or get arrested before trial.

Make a Decision

Please decide what action to take for this defendant.

○ Release the defendant.
○ Detain the defendant.

B Loans

Loan Applicant Profile

The loan applicant has applied for a loan of $5,300, with an interest rate of 14.08%. The loan will be paid in 36 monthly installments of $181.35. The applicant has an annual income of $70,000 and a Good credit score. The applicant is a home owner.

Risk Assessment Algorithm

The risk assessment algorithm predicts that this person is 20% likely to default on their loan.

Make a Predicion

How likely is this loan applicant to default on their loan?

○ 0% ○ 10% ○ 20% ○ 30% ○ 40% ○ 50% ○ 60% ○ 70% ○ 80% ○ 90% ○ 100%

Figure 5.2: Examples of the prompts presented to participants. (A) A profile presented to a decision-making participant in the pretrial setting. (B) A profile presented to a prediction-making participant in the loans setting. Both of these examples are for participants in the treatment group; participants in the control group saw the same prompt, but without the section about the risk assessment.

ticipants about each subject. Because estimates of risk can be altered by risk assessments, we calculated these averages only from participants who were subject to the same risk assessment treatment as the decision-maker. That is, $\textit{perceived.risk}_{i,r} = \text{avg}(\textit{prediction}_{i,r} | I = i, \text{R} = r)$, where i is the index of the defendant or applicant in question and r is a binary indicator for whether or not the risk assessment was shown. Thus, for instance, the perceived risk for a decision about a defendant made without the risk assessment is the average of risk predictions for that same defendant made without the risk assessment.

After making predictions or decisions, participants were asked to answer several questions about their behaviors and beliefs on a 7-point Likert scale. Because people are often unaware of how particular stimuli affected their behavior [362] and are subject to social desirability bias [163], many of the exit survey questions were phrased indirectly, using the projective viewpoint, which has been shown to yield more accurate reports of behaviors and beliefs than direct questioning [163, 164].

5.2.2 Analysis

Predictions

We evaluated the quality of each prediction using an inverted Brier score bounded between 0 (worst possible performance) and 1 (best possible performance).

We measured how participants made predictions using Bayesian linear regression (we used a Bayesian approach for consistency with the next section, where Bayesian regression enabled analysis based on posteriors; in all cases the inferences made from Bayesian and non-Bayesian regressions were almost identical). We implemented models with the brms package in R [67], which provides a high-level interface to Markov Chain Monte Carlo (MCMC) sampling for Bayesian inference using Stan [70]. In both settings we regressed the average prediction about each subject (both with and without the risk assessment) on the factors presented to participants in the narrative profiles along with interactions between those factors and whether the risk assessment was shown. To

account for repeated samples of subjects (about whom risk predictions were measured both with and without the risk assessment), the model also included random effects for the subject identity. This approach allowed us to measure the influence of subject attributes and the risk assessment on the average risk prediction about each subject.

$$\begin{aligned}\text{perceived.risk} \sim\ & \text{race} + \text{gender} + \text{age} + \text{offense.type} + \text{number.prior.arrests} \\ & + \text{number.prior.convictions} + \text{prior.failure.to.appear} + \text{show.RA} \\ & + \text{race} * \text{show.RA} + \text{gender} * \text{show.RA} + \text{age} * \text{show.RA} + \text{offense.type} * \text{show.RA} \\ & + \text{number.prior.arrests} * \text{show.RA} + \text{number.prior.convictions} * \text{show.RA} \\ & + \text{prior.failure.to.appear} * \text{show.RA} + (1|\text{subject})\end{aligned} \tag{5.1}$$

$$\begin{aligned}\text{perceived.risk} \sim\ & \text{income} + \text{fico.category} + \text{own.home} + \text{monthly.installment} \\ & + \text{interest.rate} + \text{loan.amount} + \text{loan.term} + \text{show.RA} \\ & + \text{income} * \text{show.RA} + \text{fico.category} * \text{show.RA} + \text{own.home} * \text{show.RA} \\ & + \text{monthly.installment} * \text{show.RA} + \text{interest.rate} * \text{show.RA} + \text{loan.amount} * \text{show.RA} \\ & + \text{loan.term} * \text{show.RA} + (1|\text{subject})\end{aligned} \tag{5.2}$$

We initialized models with uninformative priors and implemented sampling using 4 chains with 1000 iterations, following 1000 burn-in iterations on each chain. All coefficients in both models returned $\hat{R} = 1.00$, indicating that the chains were well-mixed and have converged to a common distribution. We estimated statistical significance from the samples using the probability of direction measure and obtaining the equivalent frequentist p-value [321, 320]. These coefficients and p-values are very similar to what is obtained by fitting these same

regressions using non-Bayesian linear regression.

Decisions

We evaluated the relationship between risk predictions and decisions using Bayesian mixed-effects logistic regression, implemented in brms [67]. We treated predictions of risk as a key input to decisions about whether to detain defendants and reject loan applications [190]. In both settings, each decision made by a participant was regressed on the average risk prediction about the subject in question, whether the risk assessment was shown, and the interaction between these two factors. To account for repeated samples, the model also included random effects for the participant identity, the subject identity, and the index (1–30) marking the participant's progress in the experiment. Because these risk predictions have already accounted for the specific attributes of each subject and because we did not directly measure each decision-making participant's estimates of risk, we did not include subject attributes within this regression formula. This formula allows us to measure decision-making as a function of risk estimate.

Recall that to avoid priming participants to focus on risk, we did not ask participants making decisions for their estimate of each subject's risk. Instead, we used the predictions made by other participants to provide an estimate of how each decision-making participant perceived the risk of each subject. Because we had participants making predictions and decisions both with and without the risk assessment, we accounted for the effect of the risk assessment on predictions by calculating average predictions made both with and without the risk assessment. Thus, for decisions made with/without the risk assessment, *perceived.risk* measures the average prediction made about the same subjects with/without the risk assessment. The *perceived.risk* measurements are based on an average of 18.13±4.00 participant predictions about each subject in each treatment (RA or no-RA), with an average standard deviation in risk predictions of 21.85±6.64 and an average standard error of 5.21±1.70 (these

values are almost identical across the two settings).

$$\begin{aligned} \text{decision} \sim \text{perceived.risk} + \text{show.RA} + \text{perceived.risk} * \text{show.RA} \\ + (1|\text{participant}) + (1|\text{subject}) + (1|\text{progress.idx}) \end{aligned} \tag{5.3}$$

If risk assessments simply present information that improves human estimates of risk (Setting 3), we would expect to see that, conditioned on a given level of perceived risk, the risk assessment does not alter decisions. In this case, both regression factors that include *show.RA* would be nonsignificant. Yet if risk assessments also influence decision-making (Setting 4), we would expect to see that people are more attentive to reducing risk when making decisions. This result could emerge through two different effects: 1) participants being more risk-averse at all levels of risk (in this case, the show.RA factor would be positive), or 2) participants being more sensitive to increases in risk (in this case, the *perceived.risk* $*$ *show.RA* factor would be positive).

We initialized models with uninformative priors and implemented sampling using 4 chains with 1000 iterations, following 1,000 burn-in iterations on each chain. In both models, all fixed effect coefficients returned $\hat{R} = 1.00$ and all random effect coefficients returned $\hat{R} \leq 1.01$, indicating that the chains were well-mixed and have converged to a common distribution. We estimated statistical significance from the samples using the probability of direction measure and obtaining the equivalent frequentist p-value [321, 320]. The coefficients and p-values are very similar to what is obtained by fitting these same regressions using standard logistic regression.

To obtain the estimated values (and standard deviations) of the fitted decision functions we took all 4,000 posterior samples of the fixed effect coefficients from the fitted model. We then used each set of coefficients to calculate the rate of detaining defendants or rejecting loan applicants at each level of risk from 0% to 100% (in intervals of 0.1%) both with and without the risk assessment. We also used these posterior estimates for the fitted decision rates to determine, at each level of risk, the shifts in negative decision rates caused by the risk assessment.

5.2.3 Simulations

Because participants in our experiments either were or were not exposed to the risk assessment, what we observed in the experiments was the results of Settings 1 and 4: people whose predictions and decisions were subject to the same stimuli. Estimating the effect of the shifts in decision-making requires disentangling the risk assessments' effects on predictions and on decisions. This means comparing Settings 3 and 4 to determine how the changes in decision-making caused by the risk assessments affect outcomes *conditioned on making predictions using the risk assessment*.

We used simulations to distinguish the effects of changes in predictions and changes in decision-making due to the risk assessments. This meant simulating outcomes in the four settings described in Table 5.1. First, we used data from the experiments to learn prediction and decision functions both with and without the risk assessments. We estimated the outcomes in all four settings described in Table 5.1 by applying these models in various combinations to a large sample of defendants and loan applicants (e.g., we estimated the results in Setting 3 by simulating predictions "with" the risk assessment and then simulating decisions "without" the risk assessment; because the decision function depends in part on predicted risk, we treat the prediction function output as an input to the decision function). We then ran 1,000 trials simulating the outcome for each subject in each of these four settings.

Fitting Prediction and Decision Models

We began by learning the prediction and decision functions that explain the average risk predictions and negative decision rates for each defendant and loan applicant. For predictions, we used Equations S1 and S2, modeling the average risk prediction about each subject based on all seven attributes of that subject that were visible to participants as well as the interactions between those attributes and whether the risk assessment was shown. We used a similar formula for decisions, in this case modeling the negative decision rate about each subject using

the same factors as in the predictions model while also adding the average risk prediction about that subject and the interaction between that prediction and whether the risk assessment was shown:

$$\begin{aligned} \text{detention.rate} \sim\ & \text{perceived.risk} + \text{race} + \text{gender} + \text{age} + \text{offense.type} + \text{number.prior.arrests} \\ & + \text{number.prior.convictions} + \text{prior.failure.to.appear} + \text{show.RA} \\ & + \text{perceived.risk} * \text{show.RA} + \text{race} * \text{show.RA} + \text{gender} * \text{show.RA} \\ & + \text{age} * \text{show.RA} + \text{offense.type} * \text{show.RA} + \text{number.prior.arrests} * \text{show.RA} \\ & + \text{number.prior.convictions} * \text{show.RA} + \text{prior.failure.to.appear} * \text{show.RA} \end{aligned} \tag{5.4}$$

$$\begin{aligned} \text{rejection.rate} \sim\ & \text{perceived.risk} + \text{income} + \text{fico.category} + \text{own.home} + \text{monthly.installment} \\ & + \text{interest.rate} + \text{loan.amount} + \text{loan.term} + \text{show.RA} \\ & + \text{perceived.risk} * \text{show.RA} + \text{income} * \text{show.RA} + \text{fico.category} * \text{show.RA} \\ & + \text{own.home} * \text{show.RA} + \text{monthly.installment} * \text{show.RA} + \text{interest.rate} * \text{show.RA} \\ & + \text{loan.amount} * \text{show.RA} + \text{loan.term} * \text{show.RA} \end{aligned} \tag{5.5}$$

We fit all models using generalized linear regression with a logit link function from the "quasibinomial" family. We use this quasibinomial approach because the fitted value of all regressions is a probability (either a risk prediction or negative decision rate that ranges from 0%-100%) rather than a binary outcome. Although linear regression yields very similar results to what is described below, it does not guarantee that predicted values on new data will be bounded [0,1].

We used leave-one-out cross validation to test the effectiveness of this approach on out-of-sample data. Recall that we had a sample of 300 subjects in each setting, with predictions/decisions about that subject both with

and without the risk assessments, for a total training set of 600 data points. We removed predictions/decisions about one subject at a time, trained the model on the data about the other 299 subjects, and estimated the prediction/decision that would be made about the held-out subject both with and without the risk assessment. In this manner we obtained out-of-sample predictions about the full set of data to evaluate. We tested the prediction and decision models independently (i.e., using the empirical average predictions as input for the decision functions) before testing the full pipelines (in which the estimated risk predictions are used as input for the decision functions).

The mean average error (MAE) on the full pipeline is 5.92 (RMSE=7.46) in the pretrial setting and 7.33 (RMSE=9.95) in the loans setting. In both settings the performance of the full pipeline decisions model is similar to that of the independent decisions model. All the models are unbiased estimators, with mean errors close to 0.

We then fit prediction and decision models for both settings on the full set of 300 subjects, for use in our simulations.

Predictions on New Subjects

We applied these models to a large, representative set of subjects that were not shown to participants in the experiments: the held-out validation sets from both settings that were described in Section 1.2 (not including the 300 subjects that were sampled for inclusion in our experiments). These samples represent approximately 10% of the full data in each setting and contain 4,375 defendants and 4,231 loan applicants drawn randomly from the populations described in Tables S1 and S2. Both of these samples are representative of the full population reflected in the datasets (recall that our 300-defendant sample was not fully representative due to privacy restrictions).

These simulations proceeded as follows:

1. Apply the predictions model to duplicates of every subject, one in which the risk assessment is coded

as not being shown and another in which the risk assessment is coded as being shown. This allows us to obtain two estimated average risk predictions about each subject (one made "with" and one made "without" the risk assessment).

2. Apply the decisions model to duplicates of every prediction about subjects, again with one decision in which the risk assessment is coded as not being shown and another in which the risk assessment is coded as being shown. For all of the predictions made "with" the risk assessment, for example, we estimated the negative decision rates if decisions were made "with" or "without" the risk assessment. This process yields four estimated negative decision rates for each subject, which are based on the four possible decision-making processes: predictions and decisions are both made without the risk assessment, predictions are made without the risk assessment but decisions are made with the risk assessment, predictions are made with the risk assessment but decisions are made without the risk assessment, and predictions and decisions are both made with the risk assessment.

3. Run 1,000 trials simulating the outcome for each subject based on the negative decision probabilities estimated in Step 2. This allowed us to estimate the distribution of outcomes for the four decision-making processes described above.

5.3 Results

5.3.1 2.1 COVID-19 Reliability Analysis

In order to ensure that any observed results would not be the effects of aberrant behavior during the COVID-19 pandemic, immediately before running our main experiments in May 2020 we conducted a retest of a trial experiment conducted in December 2019.

The December 2019 trial closely resembled the experiments described Section 1. We recruited 240 participants

from Mechanical Turk to evaluate a test sample of 100 criminal defendants. For the May 2020 trial we recruited 250 participants to evaluate the same set of 100 criminal defendants. We compared the results of these two trials in order to determine whether people's perceptions or behaviors in response to COVID-19 (or changes in the population of Mechanical Turk workers) were likely to alter the results of our experiments. We focused on three results central to our study: the demographics of participants in our experiments, the manner in which participants made predictions of risk, and the manner in which participants made decisions about whether to release or detain defendants.

Participant Demographics

The demographics of our study participants were similar across the two trials. In both cases, participants were predominantly white (80.5% in 12/19 vs. 73.4% in 05/20), male (58.6% vs. 58.0%), and college educated (73.5% vs. 70.2%). A logistic regression predicting which trial participants were part of, based on all of the demographic attributes reported during the introductory survey, yielded no terms that were statistically significant.

Prediction Function

Among participants tasked with making predictions, we observed a high degree of consistency between the predictions made across the two trials. The correlation between the average prediction made about each of the 100 defendants was $r(198)=+.94$, $p<.001$. A two-sided t-test yielded no statistically significant difference between the average prediction performance of participants across the two trials (0.751 vs. 0.753, $p=.82$).

We also estimated the function used by participants to predict the risk of each criminal defendant. Akin to our analysis of predictions described below, we used a mixed-effects linear regression model to measure the average risk prediction about each defendant, grouped based on whether or not the risk assessment was shown and whether or not the prediction was made in the first (12/2019) or second (05/2020) trial. The model included fixed effects for whether the risk assessment was shown, whether the predictions were made in the first or second

trial, the attributes of defendants, and the interactions between these three sets of factors (up to three-way). We also included a random effect for each defendant to account for the repeated predictions by each participant and about each defendant. Overall, we observed minimal differences in the effect of these attributes on predictions across the two trials. The trial number and the interaction between trial number and whether the risk assessment was presented were not statistically significant. Only two of the interactions that included trial number were statistically significant, as participants were slightly less responsive to prior failures to appear (P=.025) and prior convictions (P=.039) in the second trial.

Decision Function

We also observed a high degree of consistency between the two trials among participants tasked with making release/detain decisions about criminal defendants. The correlation between the average detention rate for each of the 100 defendants was $r(198)=+.97$, $p<.001$.

We also estimated the function used by participants to decide whether to release or detain each criminal defendant. Akin to the primary analysis of decisions described below, we used a mixed-effects logistic regression model on all 8,070 decisions made across the two trials. The model included fixed effects for whether the risk assessment was shown, the trial number, and the average prediction of risk about each defendant (in the applicable treatment and trial number), with up to three-way interactions between these factors. We included random effects for participants, defendants, and status in the experiment to account for repeated measurements. None of the coefficients that included trial number were statistically significant, indicating that decision-making did not notably differ across the December 2019 or the May 2020 trials.

Summary

In sum, we find high levels of test-retest reliability: the results found in May 2020 (in the midst of the COVID-19 pandemic) closely resembled the results found in December 2019, suggesting that our results are not merely the

product of, nor notably influenced by, aberrant behaviors that arose in response to COVD-19. These results—which indicate a high degree of consistency in Mechanical Turk participant predictions and decisions across experiments separated by approximately 4.5 months—also indicate the reliability of our results more generally as being reproducible upon repeated experimentation.

5.3.2 Participants

We conducted trials on Mechanical Turk over the course of two weeks in May 2020. 2,685 participants completed the experiments. Filtering out data from workers who failed at least one of the attention check questions in the intro and exit surveys or who required more than four attempts to pass the comprehension test yielded 2,140 participants for our analysis.

The participant population is described in Table 5.4. Across both settings, a majority of participants were male, white, and have completed at least a college degree. Measures of familiarity with certain topics, clarity of the experiment, and how enjoyable the experiment was to complete are based on participant self-reports measured on a Likert scale from 1 (low) to 7 (high).

5.3.3 Effect of Risk Assessments on Predictions

Presenting participants with the risk assessment improved prediction accuracy and reduced risk estimates (Figure 5.3). In the pretrial setting, the average participant prediction quality increased from 0.72 to 0.75 ($P<.001$, $d=0.11$). A paired t-test comparing the average predictions of risk about each defendant finds that the risk assessment reduced perceived risk by an average of 1.6% about each defendant (from an average 40.6% to 38.9%, $P=.001$, $d=0.19$). While the reduction in perceived risk was significant for white defendants (38.4% to 35.7%, $P=.003$, $d=0.30$), Black defendants received a smaller and nonsignificant reduction (41.7% to 40.7%, $P=.085$, $d=0.12$). The improvement in prediction quality from showing the risk assessment was larger in the loans setting, from 0.75 to 0.83 ($P<.001$, $d=0.31$). The risk assessment also altered predictions of risk more dramatically in

	Pretrial N=1,040	Loans N=1,100
Demographics		
Male	59.8%	61.0%
Black	14.2%	11.9%
White	71.5%	72.9%
18-24 years old	7.4%	6.8%
25-34 years old	46.1%	45.0%
35-59 years old	43.0%	43.9%
60+ years old	3.6%	4.3%
College degree or higher	82.5%	81.9%
Criminal justice familiarity	5.1	5.1
Financial lending familiarity	4.9	5.1
Machine learning familiarity	4.7	4.8
Treatment		
Decisions, no RA	39.2%	38.9%
Decisions, with RA	35.0%	36.0%
Predictions, no RA	13.3%	13.2%
Predictions, with RA	12.5%	11.9%
Outcomes		
Average hourly wage	$14.86	$15.16
Experiment clarity	6.4	6.4
Participant enjoyment	5.8	5.9

Table 5.4: Attributes of the participants in our experiments, by setting.

the loans setting, reducing the perceived risk for 92.3% of loan applicants, with an overall average reduction of 14.2% (from 38.5% to 24.3%, P<.001, d=1.54). These results are consistent with the prior two chapters.

The risk assessment also altered decisions in both settings, albeit in different directions. A paired t-test finds that the risk assessment reduced each defendant's likelihood of pretrial detention by an average of 2.4% (from an average of 44.5% to 42.1%, P<.001, d=0.21). White defendants received a slightly larger average reduction (38.7% to 35.9%, P=.014, d=0.24) than Black defendants (47.7% to 45.5%, P=.007, d=0.20). In the loans setting, despite prompting significant reductions in risk predictions, the risk assessment nonsignificantly increased loan rejection rates by an average of 1.0% per subject (from 22.1% to 23.1%, P=.159, d=0.08). This pattern of the risk assessment reducing estimates of default risk but potentially increasing rejection rates in the loans setting

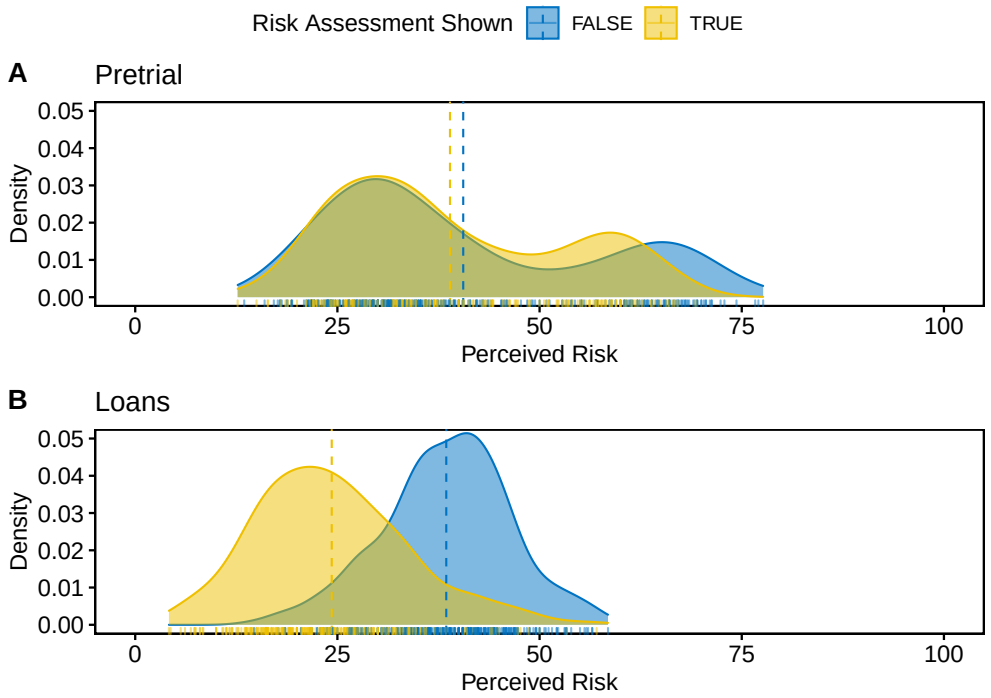

Figure 5.3: Distributions of perceived risk about subjects, by risk assessment treatment. (A) Distributions in the pretrial setting. (B) Distributions in the loans setting. Dotted lines indicate the average value in each treatment. The risk assessment caused predictions of risk to decrease for 54.0% of defendants (59.8% of whites and 50.8% of Blacks) and 92.3% of loan applicants.

indicates that reducing risk predictions does not directly translate to equivalent changed in decisions. Instead, decisions are relatively inelastic to shifts in predictions in both settings: for instance, a 10% reduction in perceived risk is associated with a 4.4% reduction in the pretrial detention rate and a 2.8% increase in the loan rejection rate (Figure 5.4).

In both settings, the risk assessment improved prediction accuracy by by aligning people's predictions more closely with those of the risk assessments, prompting participants to adjust the risk they associated with certain factors and to more strongly account for factors that participants without the risk assessment ignored (Table 5.5 and Table 5.6). In the pretrial setting, predictions of risk without the risk assessment were influenced primarily by the type of crime for which the defendant was arrested, the defendant's number of prior arrests (+0.72% risk for each arrest), and whether the defendant had a previous failure to appear (+27.82%). The risk assessment induced several shifts, most notably increasing the baseline prediction (+6.98%), prompting participants to consider the

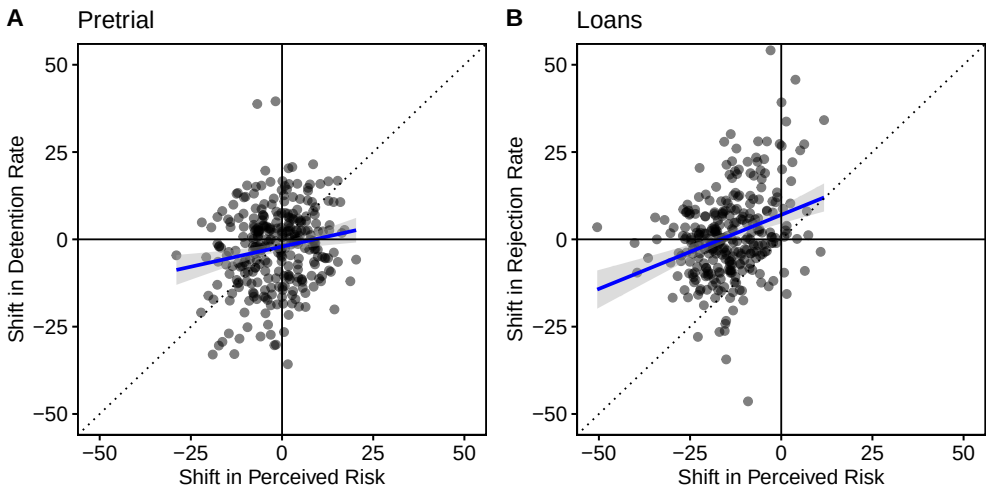

Figure 5.4: Shifts in predicted risk and negative decision rates for each subject caused by showing the risk assessment to participants. (A) Pretrial setting. Results for Black and white defendants are pooled because they are very similar. (B) Loans setting. Positive values on the x-axis indicate that the risk assessment increased the average risk prediction about a subject. Positive values on the y-axis indicate that the risk assessment increased the detention or rejection rate about a subject. The blue lines indicate linear regression fits of decision shifts versus prediction shifts. The intercept is negative in the pretrial setting (−2.07, P=.002) and positive in the loans setting (7.02, P<.001). The coefficients on prediction shifts are less than 1 (0.23 in pretrial, P=.003; 0.42 in loans, P<.001), indicating that decisions are relatively inelastic to shifts in predictions.

age of defendants (–0.20% risk for each year of age), and reducing the risk associated with prior failures to appear (–7.43%). In the loans setting, predictions of risk without the risk assessment were influenced primarily by the applicant's annual income (–0.03% risk for every $1,000) and FICO score as well as the loan's interest rate (+0.33% for each percent interest). The risk assessment significantly reduced participants' baseline prediction (–24.02%), increased the salience of annual income (–0.02%) and interest rate (+0.50%), and prompted participants to consider the length of the loan (+7.41% risk for a 60-month term). These shifts in both settings brought the human predictions closer in line with how the risk assessment made predictions. Although the risk assessment did improve prediction accuracy, people collaborating with the risk assessment underperformed the risk assessment alone in both settings (P<.001).

	Not Shown RA	Shown RA (interaction)
Intercept	27.88 (1.50) ***	+6.98 (2.03) ***
White	–0.03 (0.72)	–0.98 (0.98)
Male	0.04 (0.91)	–0.42 (1.25)
Age	0.03 (0.04)	–0.20 (0.05) ***
Property crime	–2.29 (0.74) ***	+0.43 (1.04)
Public order crime	–0.28 (1.59)	–3.50 (2.21)
Violent crime	3.00 (0.95) ***	–7.45 (1.27) ***
Number of prior arrests	0.72 (0.17) ***	+0.20 (0.23)
Number of prior convictions	0.31 (0.17) .	+0.09 (0.22)
Prior failure to appear	27.82 (1.33) ***	–7.43 (1.76) ***

Table 5.5: Bayesian linear regression results estimating the average risk prediction about each defendant. Regressions are based on the attributes of each defendant, whether the risk assessment was shown, and interactions between these factors. The first column presents the coefficient of each factor and the second column presents the interaction of that factor with the risk assessment. The shifts in prediction-making indicated here brought participant predictions closer in line with how the risk assessment made predictions. Parenthetical terms represent standard errors. . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

5.3.4 Effect of Risk Assessments on Decisions

Our Bayesian mixed-effects logistic regressions indicate that the risk assessment increased participant attentiveness to risk when making decisions in both settings, thus demonstrating that risk assessments prompt a shift to Setting 4 rather than Setting 3. In the pretrial setting, the risk assessment altered decisions by making participants more sensitive to increases in risk (Figure 5.5). This means that perceived risk more strongly influenced whether defendants were released or detained: the risk assessment reduced pretrial detention rates for low levels of perceived risk but increased pretrial detention rates for high levels of perceived risk. While a 10% increase in perceived risk increased the odds of pretrial detention by a factor of 1.82 without the risk assessment, for participants shown the risk assessment a 10% increase in perceived risk increased the odds of detention by a factor of 2.39 (Table 5.7). Thus, for example, an increase in perceived risk from 30% and 60% led to an increase in detention probability of 42.0% without the risk assessment and of 57.0% with the risk assessment (Table 5.9).

In the loans setting, the risk assessment altered decisions by making participants more risk-averse at all levels of risk (Figure 5.5). Presenting the risk assessment increased the odds of rejecting loan applications by a factor

	Not Shown RA	Shown RA (interaction)
Intercept	39.37 (1.93) ***	–24.02 (2.45) ***
Annual income	–0.03 (0.01) ***	–0.02 (0.01) *
Good FICO score	–5.81 (1.04) ***	+2.23 (1.31) .
Very good FICO score	–7.91 (1.46) ***	+1.47 (1.83)
Exceptional FICO score	–9.29 (2.52) ***	–0.51 (3.24)
Fully own home	–0.30 (0.99)	+2.13 (1.26) .
Loan amount	0.27 (0.28)	–0.45 (0.37)
Monthly installment	–0.74 (8.96)	+16.81 (11.50)
Interest rate	0.33 (0.12) **	+0.51 (0.15) ***
60-month term	–2.21 (1.91)	+7.41 (2.49) **

Table 5.6: Bayesian linear regression results estimating the average risk prediction about each loan applicant. Regressions are based on the attributes of each application, whether the risk assessment was shown, and interactions between these factors. Annual income, loan amount, and monthly installment are measured in units of $1000. The first column presents the coefficient of each factor and the second column presents the interaction of that factor with the risk assessment. The shifts in prediction-making indicated here brought participant predictions closer in line with how the risk assessment made predictions. Parenthetical terms represent standard errors.. p<0.1; * p<0.05; ** p<0.01; *** p<0.001

	Not Shown RA	Shown RA (interaction)
Intercept	–2.74 (0.17) ***	–1.14 (0.14) [0.32] ***
Predicted risk	0.60 (0.04) [1.82] ***	+0.27 (0.03) [1.31] ***

Table 5.7: Bayesian mixed-effects logistic regression results estimating the average risk prediction about each defendant. Regressions are based on the average predicted risk about the defendant, whether the risk assessment was shown, and interactions between these factors. Presenting the risk assessment increased participants' sensitivity to increases in risk, reducing the likelihood of detention for 0% risk but increasing the rate at which detention probability increases as predicted risk increases. Risk is measured in units of 10%. The first column presents the coefficient of each factor and the second column presents the interaction of that factor with the risk assessment. Parenthetical terms represent standard errors. Terms in brackets represent odds ratios. The standard deviation for the random effects are 1.03 for worker, 0.90 for subject, and 0.07 for experiment progress index. . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

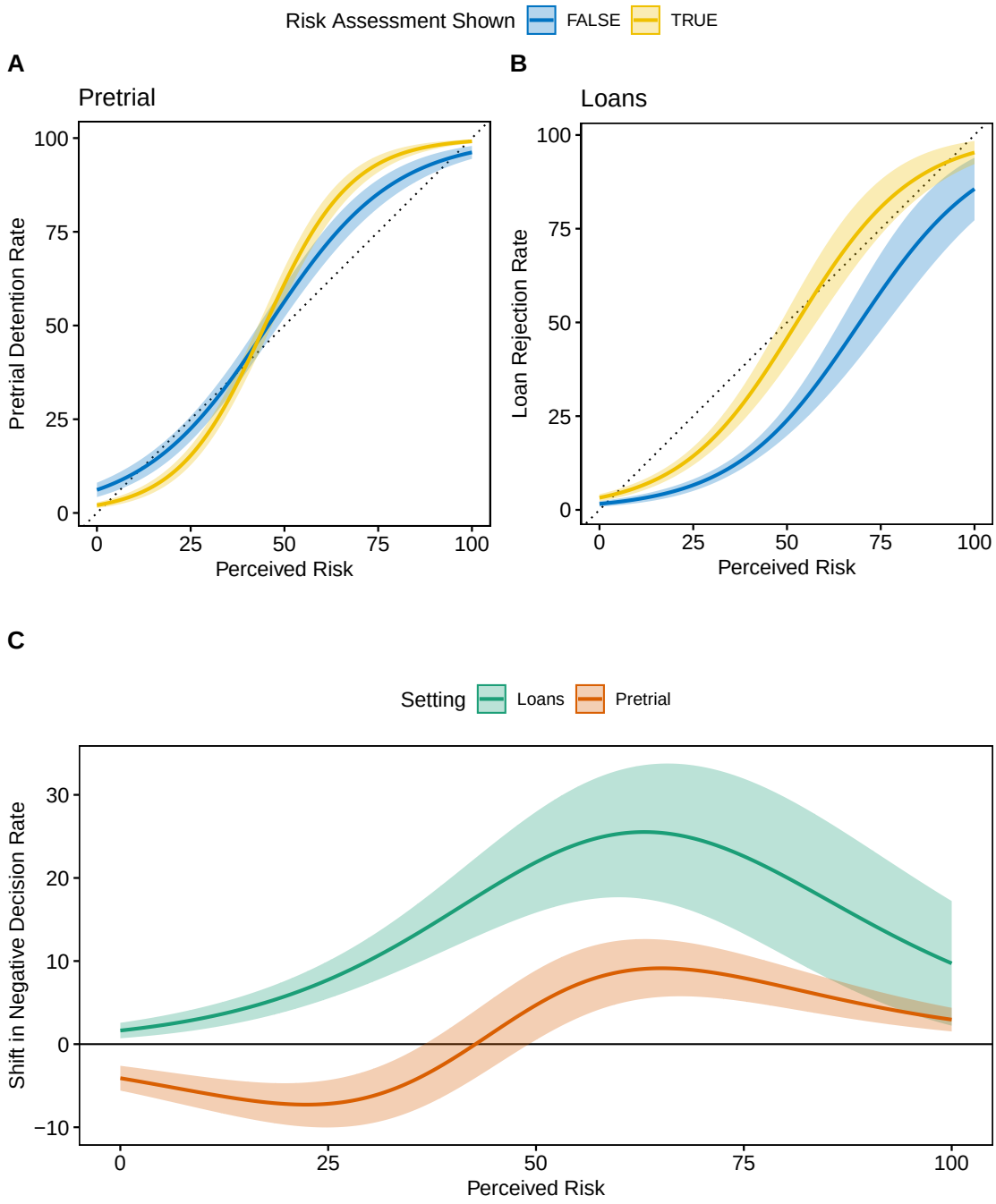

Figure 5.5: Change in decision-making caused by showing the risk assessment to participants. (A) Decision functions indicating the likelihood of detaining a defendant as a function of the perceived risk about that defendant, by risk assessment treatment. The risk assessment makes people more sensitive to increases in risk, reducing detention at low risk and increasing detention at high risk. (B) Decision functions indicating the likelihood of rejecting a loan application as a function of the perceived risk about that applicant, by risk assessment treatment. The risk assessment causes rejection rates to increase at all levels of risk. (C) Shift in negative decision (pretrial detention or loan rejection) rate due to the shift in decision-making caused by showing the risk assessment, by setting. For instance, given a perceived risk of 50% the risk assessment increases the likelihood of pretrial detention by 4.7% and the likelihood of loan rejection by 21.9%. Bands indicate 95% confidence intervals all in panels.

	Not Shown RA	Shown RA (interaction)
Intercept	–4.15 (0.24) ***	+0.74 (0.22) [2.09] ***
Predicted risk	0.60 (0.05) [1.82] ***	+0.05 (0.05) [1.05]

Table 5.8: Bayesian mixed-effects logistic regression results estimating the average risk prediction about each loan applicant. Regressions are based on the average predicted risk about the loan applicant, whether the risk assessment was shown, and interactions between these factors. Presenting the risk assessment increased the odds of rejecting loan applications by a factor of 2.09 but did not affect participants' sensitivity to increases in risk. Risk is measured in units of 10%. The first column presents the coefficient of each factor and the second column presents the interaction of that factor with the risk assessment. Parenthetical terms represent standard errors. Terms in brackets represent odds ratios. The standard deviation for the random effects are 1.19 for worker, 0.90 for subject, and 0.29 for experiment progress index. . p<0.1; * p<0.05; ** p<0.01; *** p<0.001

of 2.09 (Table 5.8). For all levels of perceived risk up to 46.0% (covering 97.3% of risk estimates with the risk assessment), participants were more than twice as likely to reject loan applications if they were shown the risk assessment (Table 5.9). For instance, an applicant with a perceived risk of 30% had an 8.7% likelihood to be rejected by a participant not shown the risk assessment but an 18.8% likelihood to be rejected by a participant shown the risk assessment.

Participant responses to survey questions after making decisions indicate that shifts in people's decisions did not align with shifts in their perceptions or beliefs. Despite becoming more attentive to risk when making decisions, participants presented with the risk assessment expressed less support for basing decisions on risk (Pretrial: P=.003, d=0.21; Loans: P=.001, d=0.23) and did not alter the priority given to key considerations (including risk) when making decisions (Table 5.10).

5.3.5 Distinguishing Between Shifts in Predictions and Decisions

Because the risk assessment influenced both prediction-making and decision-making, our results reflect the behavior of participants whose predictions and decisions were both subject to the same stimuli (i.e., Settings 1 and 4). Because we did not observe the outcomes of Setting 3, we cannot directly compare Settings 3 and 4 in order to isolate the effects of the risk assessment's unexpected influence on decision-making, conditioned on the

Risk	No RA	**Pretrial** Shown RA	Difference	No RA	**Loans** Shown RA	Difference
0%	6.15%	2.06%	–4.09% [5.38]	1.60%	3.24%	+1.64% [3.45]
10%	10.62%	4.74%	–5.88% [5.93]	2.84%	5.98%	+3.15% [4.65]
20%	17.73%	10.52%	–7.21% [5.64]	5.00%	10.82%	+5.82% [6.10]
30%	28.13%	21.80%	–6.33% [3.84]	8.70%	18.81%	+10.11% [7.17]
40%	41.58%	39.83%	–1.75% [0.88]	14.73%	30.68%	+15.95% [7.34]
50%	56.41%	61.11%	+4.70% [2.20]	23.89%	45.78%	+21.89% [7.07]
60%	70.16%	78.83%	+8.67% [4.49]	36.31%	61.63%	+25.32% [6.49]
70%	81.01%	89.80%	+8.79% [5.52]	50.80%	75.27%	+24.47% [5.39]
80%	88.54%	95.41%	+6.87% [5.35]	65.04%	85.19%	+20.14% [4.14]
90%	93.32%	98.00%	+4.68% [4.71]	76.93%	91.55%	+14.63% [3.19]
100%	96.19%	99.14%	+2.95% [4.07]	85.60%	95.32%	+9.72% [2.54]

Table 5.9: Negative decision probabilities at a range of risk levels, by setting and risk assessment treatment. No RA indicates the probability of negative decisions when not shown the risk assessment, Shown RA indicates the probability of negative decisions when shown the risk assessment, and Difference indicates the difference between these values (numbers in brackets indicate the effect size of this difference). All differences in both settings are statistically significant with p<.001.

	Not Shown RA	Shown RA	P-value	Effect size
Pretrial				
Incapacitation	30.86	29.89	.341	0.07
Freedom	25.76	26.68	.372	0.07
Deterrence	20.04	19.05	.245	0.08
Rehabilitation	23.35	24.38	.289	0.08
Loans				
Likelihood to pay	40.98	39.28	.211	0.09
Equity	21.51	22.59	.124	0.11
Economic development	19.63	19.29	.622	0.03
Neighborhood stability	17.89	18.84	.200	0.09

Table 5.10: Participant beliefs about how decision-makers should balance priorities. After making decisions, participants were asked to what extent a decision-maker (a judge or government loan agent) should value four salient considerations when making decisions. Participants had to assign a total of 100 points (in increments of 5) across the four considerations. None of the average values assigned to these considerations differ significantly across the risk assessment treatment.

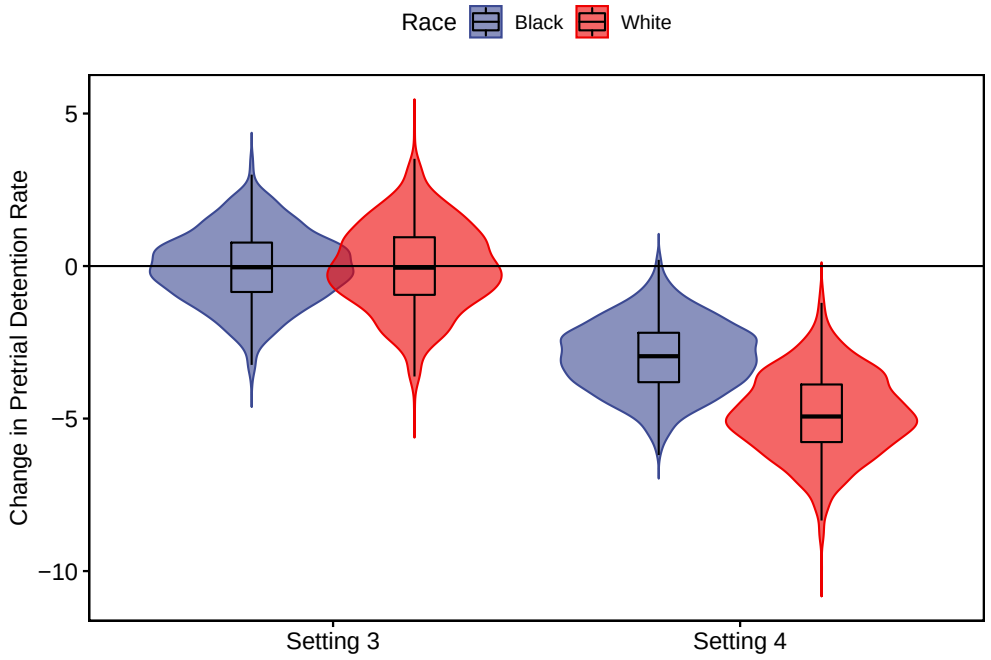

Figure 5.6: Simulated changes in pretrial detention rates in Settings 3 and 4 compared to Setting 1, by race. In Setting 3, the detention rate for both races is reduced by less than 0.1% compared to Setting 1. In Setting 4, the detention rate for Black defendants is reduced by 3.0% while the detention rate for white defendants is reduced by 4.9% compared to Setting 1.

expected effect of the risk assessment improving predictions. We estimated this effect by simulating predictions and decisions about more than 4,000 defendants and loan applicants.

In the pretrial setting, the risk assessment's influence on decision-making reduced the average detention rate but exacerbated racial disparities, an effect also observed in empirical studies of pretrial risk assessments [8, 107, 464, 466]. Compared to the baseline process in Setting 1, the risk assessment's effect on predictions alone (Setting 3) did not alter detention rates for either race whereas the risk assessment's effect on predictions and decisions (Setting 4) reduced detention by 4.9% for white defendants and 3.0% for Black defendants ($P<.001$, $d=1.52$; Figure 5.6). Thus, the shift in decision-making prompted by the risk assessment increased the racial disparity by 1.9% and by a factor of 1.34 from 5.6% in Setting 3 to 7.5% in Setting 4 ($P<.001$, $d=1.06$; Figure 5.7).

In the loans setting, the change in decision-making caused by the risk assessment generated a marked increase in rejections. Were the risk assessment to affect only predictions, the simulated rejection rate would drop from

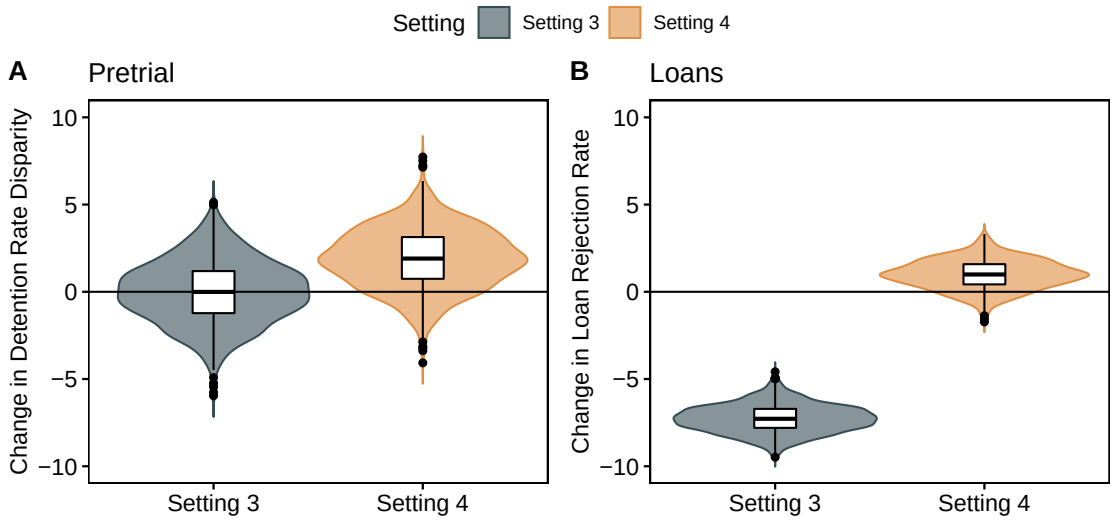

Figure 5.7: Simulated changes in outcomes in Settings 3 and 4 compared to Setting 1. (A) Change in Black-white detention rate disparity in Settings 3 and 4 compared to Setting 1. Setting 3 reduced the average racial disparity by less than 0.1% while Setting 4 increased the average racial disparity by 1.9%. (B) Change in loan rejection rate in Settings 3 and 4 compared to Setting 1. Setting 3 reduced the average rejection rate by 7.3% while Setting 4 increased the average rejection rate by 1.0%. Comparisons of Settings 3 and 4 measure the effect of the risk assessment's influence on decision-making, conditioned on the expected effect of the risk assessment improving predictions.

22.2% in Setting 1 to 14.9% in Setting 3 (P<.001, d=13.09). The shift in decision-making negates this effect, however, increasing rejection rates by 8.3% and by a factor of 1.55 from 14.9% in Setting 3 to 23.2% in Setting 4 (P<.001, d=14.88). Instead of simply reducing predictions of risk and thereby generating a 7.3% increase in loans granted, the risk assessment also increased risk-aversion and thereby reduced the loans provided by 1.0% (Figure 5.7).

5.4 Alternative Explanations

In this section we discuss potential alternative explanations for our findings (in contrast to the explanation that the risk assessment makes risk a more salient factor in decision-making) and describe why they are inconsistent with our experimental results.

5.4.1 Participants Have Greater Confidence in Risk Predictions

One alternative explanation is that the risk assessment makes people more confident in their risk prediction rather than more concerned about avoiding risk in decision-making. In other words, people may place a greater weight on their risk prediction because they are more certain about this prediction (rather than because they are more concerned about risk as a consideration). If this were the case, we would expect to see risk becoming a more "extreme" distinguishing factor in decisions: low levels of risk have even lower detention/rejection rates, while high levels of risk lead to higher rates. That is indeed what we observe in the pretrial setting (Figure 5.5), meaning that the results appear consistent with both our explanation as well as this alternative explanation. We observe a quite different effect in the loans setting, however: rejection rates go up at all levels of risk (Figure 5.5). This result is consistent with our explanation that the risk assessment makes people more risk-averse yet inconsistent with people becoming more confident in their risk prediction. For instance, it is relatively implausible that becoming more confident that a loan applicant has a 10% likelihood to default on the loan would more than double the likelihood of rejecting that loan application. Thus, the loans setting results are consistent with our explanation of risk-aversion but inconsistent with the alternative explanation of greater confidence.

We can look to participant self-reports of confidence to further investigate the role of confidence in decision-making, finding that *the risk assessment has no significant effects on participant confidence.* In the exit survey at the end of the experiment, every participant was asked how confident they were in their decisions, on a Likert scale from 1 (least confident) to 7 (most confident). Across both predictions and decisions in both settings, the risk assessment did not affect participant confidence. In the pretrial setting, participants making predictions reported an almost identical average confidence of 5.30 both with and without the risk assessment (P=.978, d=0.00). Participants making decisions reported being negligibly more confident (P=.246, d=0.08). In the loans setting, the risk assessment negligibly increased participant confidence among participants making predictions (P=.580, d=0.07) and decisions (P=.213, d=0.09). Given that the risk assessment did not produce any significant

impacts on participant self-reports of confidence, it seems quite unlikely that the effects of the risk assessment can be attributed to participants being more confident in their estimates of risk when making decisions.

Finally, even if the alternative explanation does hold in the pretrial setting, the ultimate effects are the same. Whether because people are more confident in their risk prediction or because they are more concerned about risk, the result is that risk becomes a more important factor distinguishing between who is detained and who is released before trial. This represents a substantial and unexpected change in policy toward more strongly making pretrial decisions on the basis of risk, a shift that has been heavily debated for decades.

5.4.2 Prediction-Makers and Decision-Makers Have Different Predictions of Risk

Another alternative explanation is that perceived risk differs between people making predictions and people making decisions. Recall that in our experiments, we estimated the perceived risk for decision-makers by taking the average perceived risk about the same subject from predictors (controlling for whether the risk assessment shown to each group). It is plausible, however, that these two groups do not have identical perceptions: in particular, the effect of the risk assessment on predictions may be attenuated for participants who were not explicitly asked to report a prediction. Because decision-makers were not asked to make an explicit estimate of risk, these participants may not have had their internal estimate of risk be as strongly influenced by the risk assessment. Although it is possible that decision-makers and predictors do not share identical perceptions of risk, this explanation is directly contradicted by some of our results. Most notable is the contrast between the effects of the risk assessment in the loans setting, reducing predictions of risk yet increasing loan rejections. As described above, the risk assessment reduced the average prediction of loan default risk from 38.5% to 24.4% ($P<.001$, $d=0.59$) and caused predictions of risk to decrease for 92.3% of loan applicants. Despite this, the risk assessment increased the loan rejection rate from 22.0% to 23.3% ($P=.016$, $h=0.03$) and caused loan rejections to increase for 50.0% of applicants (including 47.3% of the loan applicants whose perceived risk was reduced by

the risk assessment). This contrast between the effects of the risk assessment on predictions and on decisions is clearly inconsistent with decision-makers simply experiencing a diminished shift in risk perceptions compared to predictors due to the risk assessment.

5.4.3 The Risk Assessment Provides a Random Shock to Decisions

A third alternative explanation is that the risk assessment provides a random shock to decision-making, adding "noise" to decisions in a manner that is not connected to perceived risk (or changes in perceived risk). Two results can most clearly rule out this explanation. First, we observed that the reduction in pretrial detention and the increase in loan rejections were statistically significant, indicating that the risk assessment does influence decisions in specific directions (although that direction differs across settings). Second, in both settings there is a positive and statistically significant relationship between changes in perceived risk and changes in negative decision rates for each subject (Figure 5.4). Although this relationship is relatively inelastic, changes in negative decisions are significantly related to changes in perceived risk, indicating that the risk assessment's effect on decisions is connected to the risk assessment's effect on predictions.

5.5 Discussion

These results raise new concerns regarding the desirability of integrating risk assessments into government decision-making. Our evidence of risk assessments altering decision-making demonstrates that the effects of risk assessments are akin to significant shifts in policy and jurisprudence. Although risk assessments are commonly defended as merely providing information to human decision-makers, our findings demonstrate that risk assessments alter how decisions are made as a function of risk predictions, increasing the priority placed on reducing risk in pretrial detention and government loan decisions (alternative explanations, such as the risk assessment making participants more confident in their risk estimates, can be ruled out by our data and are discussed in detail

in the supplementary materials). Not only would this shift in decision-making occur without public deliberation (for it was neither intended nor expected), but it may be further shrouded by decision-makers not recognizing how the risk assessment influenced their behavior, an effect observed here as well as in the prior two chapters. If risk assessments increase the weight that judges place on risk to distinguish whom to release and detain, these algorithms would enhance the constitutionally contested policy of preventative detention [26, 326] without this effect being subject to any democratic deliberation or oversight. Similarly, greater risk-aversion in providing government loans would reduce government aid and, given that non-whites have disproportionately higher risk levels [274], potentially increase racial disparities in access to resources.

In light of these shifts in decision-making, our results demonstrate the potential harms of centering a risk-prediction framework in complex government decisions. Despite the enthusiasm for using machine learning to solve "prediction policy problems" [276], government decisions require balancing accurate predictions with numerous other values. For instance, pretrial decisions must consider the liberty of defendants [12] and government home improvement loans aim to promote equity by supporting low-income applicants [127]. Thus, even though risk assessments can improve the accuracy of risk predictions, the normative multiplicity inherent in many government decisions creates substantive conflicts between risk-minimization and other values. Studies of risk assessments may therefore overestimate the benefits and underestimate the harms of incorporating algorithmic predictions into government decisions [466, 518]. If risk assessments are to be implemented at all, they must first be grounded in rigorous evidence demonstrating what impacts they are likely to generate and in democratic deliberation supporting those impacts.

Part II

Risk and Response

Chapter 6

Predictions and Policing

6.1 Predictive Policing

Predictive policing algorithms represent another form of risk assessment used in the criminal justice system to improve public policy. These algorithms take two primary forms: place-based algorithms that predict the risk of crime in particular geographic locations and person-based algorithms that predict the risk that specific individuals will be perpetrators or victims of violence [241]. For instance, one of the most widely used place-based predictive policing tools is PredPol: software that, on the basis of historical crime records, analyzes how crime spreads between places and then forecasts that spatial process into the future to predict where the next crimes will occur. The company translates these predictions for police via an interactive map overlaid with red squares (covering 500 feet by 500 feet) at the predicted high-crime locations. If police spend time in those regions, the company posits, then they will be more effective at preventing crime and catching criminals. PredPol has aggressively shared case studies asserting the effectiveness of its software, citing "a proven track record of crime reduction in communities that have deployed PredPol" [401].

As explained by Andrew Ferguson, a legal scholar and the author of *The Rise of Big Data Policing*, predictive policing is alluring to police departments because it provides "'an answer' that seems to be removed from the hot button tensions of race and the racial tension arising from all too human policing techniques." He adds, "A black-box futuristic answer is a lot easier than trying to address generations of economic and social neglect, gang violence, and a large-scale underfunding of the educational system" [160].

Thus, in the wake of growing outrage about discriminatory police practices—including numerous high-profile police killings of African Americans—and burgeoning support for systemic police reforms, predictive policing was hailed as "a brilliantly smart idea" that could "stop crime before it starts" through objective, scientific assessments [425, 476]. In an interview, a former police analyst who served for several years as a lobbyist for the company declared, "It kind of sounds like science fiction, but it's more like science fact" [45, 402].

Thorough evaluations of predictive policing tools suggest that they promise far more than they can deliver. A 2016 study "found little evidence that today's systems live up to their claims," instead concluding, "Predictive policing is a marketing term" [419]. In fact, many of the statistics touted by PredPol are cherry-picked numbers that take advantage of normal fluctuations in crime to suggest that PredPol generated significant reductions [113]. As one statistician notes, this type of analysis "means nothing" [45].

John Hollywood, a researcher at the RAND Corporation who has assessed numerous predictive policing tools, calls any benefits of predictive policing "incremental at best" and says that to predict specific crimes "we would need to improve the precision of our predictions by a factor of 1000" [241]. Hollywood's analysis of a predictive policing effort in Louisiana—one of the only independent analyses of predictive policing that has been conducted—found that the program had "no statistically significant impact" on crime [238].

Supporters of predictive policing assert that the software must be fair, because it relies on data and algorithms. According to Brett Goldstein, Chicago's former chief data officer, an early predictive policing effort in Chicago "had absolutely nothing to do with race," because the predictions were based on "multi-variable equations" [476]. Los Angeles Police Commander Sean Malinowski called PredPol "objective" because it relies on data

[257]. Similarly, the director of Hitachi's crime-mapping software declared that the program "doesn't look at race. It just looks at facts" [455].

But the "facts" of the matter—in this case, crime statistics—are well known to be "poor measures of true levels of crime," writes the criminologist Carl Klockars. Because "police exercise an extraordinary degree of discretion in deciding what to report as crimes," Klockars explains, police statistics "are reflective of the level of police agency resources dedicated to [the] detection" of particular types of crime, rather than the actual levels of crime across society [280]. In other words, what appear to be facts about crime are largely facts about police activity and priorities.

For years, police have disproportionately targeted urban minority communities for surveillance and arrests, leading to decades of crime data that reflect this discriminatory treatment [10]. Police predominantly patrol black neighborhoods and possess significant discretion regarding when and why to arrest someone [351]. Many incidents that police never observe, act on, or even target in white communities are recorded as crimes in black neighborhoods [414].

This is what makes *The New Inquiry*'s "White Collar Crime Early Warning System" such a wonderful piece of satire. The magazine developed a model, using similar technical approaches as predictive policing tools, that predicts where financial crimes are likely to occur [294]. In Chicago, for example, whereas most crime maps show hot spots on the predominantly black and brown south and west sides, the hot spots for white-collar crime are in the central business district ("The Loop") and the primarily white north side. That these maps—and in fact the very idea of using algorithms to proactively target financial crimes—are so striking brings to light an oft-overlooked aspect of the criminal justice system and machine learning–based reform efforts: our very selection of the crimes that ought to be aggressively monitored and enforced rests in part on racist and classist notions of social order [63].

Thus, even if a machine learning algorithm is not hard-coded to exhibit racial bias, the data from which it learns reflects social and institutional biases. In this way predictive policing, while supposedly neutral, overemphasizes

the criminality of black neighborhoods and intensifies the police presence around people and places that are already unfairly targeted. An analysis in Oakland by the Human Rights Data Analysis Group demonstrates how predictive policing can lead to these disparities. Although local public health estimates suggest that drug crimes are ubiquitous in Oakland, the study found that "drug arrests tend to only occur in very specific locations—the police data appear to disproportionately represent crimes committed in areas with higher populations of non-white and low-income residents" [313]. The study's authors developed an algorithm, based on PredPol's methods, to determine what impacts predictive policing could have. They concluded that if the Oakland Police had used PredPol, "targeted policing would have been dispatched almost exclusively to lower income, minority neighborhoods" [312].

But perhaps some technical mechanisms can be employed to avoid biased predictions: if the biases reflected in crime data make predictive policing discriminatory, is there any way to make it fair? Yet the issue with predictive policing is not just that the predictions may be biased—it is that predictive policing relies on traditional definitions of crime and assumes that policing represents the proper method to address it. Focusing on the models' technical specifications (such as accuracy and bias) overlooks an even more important consideration: the policies and practices that the algorithm supports. In this way, attempts to improve social structures with mere technical enhancements implicitly (and perhaps unintentionally) subvert opportunities to critically assess and systematically reform political institutions. For even if police are dispatched to neighborhoods in the most fair and race-neutral possible manner, their typical actions once there—suspicion, stop-and-frisks, arrests—are inextricably tied to the biased practices that predictive policing was largely designed to redress. When unjust policies and practices are followed, even a superficially "fair" approach will have discriminatory impacts.

Consider what happened in Shreveport, Louisiana, during a predictive policing trial studied by RAND. When patrolling neighborhoods identified as high-crime, many police officers unexpectedly changed their tactics to focus on "intelligence gathering through leveraging low-level offenders and offenses." Officers increasingly stopped people whom they observed "committing ordinance violations or otherwise acting suspiciously" in

order to check their criminal records. Those whose histories contained prior convictions were arrested [238].

Whether or not Shreveport's algorithm accurately and fairly identified where crime would occur, it generated increased police activity and suspicion in the regions of interest. Although unintended, this response is not surprising. After all, the point of predictive policing is to identify locations where crime will occur. Doing so primes police to be "hyper alert" when patrolling inside the regions and thus to treat everyone there as a potential criminal [376]. And given the substantial evidence of racial bias in practices such as stop-and-frisk [181], it is not hard to imagine that the people whom police stop for committing violations or acting suspiciously will mainly be young men of color, thereby increasing both incarceration rates and conflict between police and communities.

Here we see the interplay between predictions and politics: whether or not predictive policing algorithms accurately and fairly identify high-crime locations, they do not dictate what actions to take in response. Governments *choose* to give responsibility for dealing with most forms of social disorder to the police. Police *choose* to go into these neighborhoods with heightened suspicion and a warrior mind-set. Thus, the seemingly technical decisions about how to develop and use an algorithm are necessarily intertwined with the clearly political decisions about the kind of society we want to inhabit. If cities truly know where crime will occur, why not work with that community and with potential victims to improve those neighborhoods with social services? Why is the only response to send in police to observe the crime and punish the offenders?

While the criminal justice system has always involved contentious and complex political decisions, the danger of using technology to make these decisions is that we will misinterpret them as technical problems that do not require political deliberation. Treating technology as the only variable of change blinds us to the full possibilities to reform the policies and practices that technology purports to improve. When predictive policing gets hailed as the new and scientific approach to policing, it distracts us from the hard choices that must be made about what police should prioritize and what their role in society should be. Thus, writes Andrew Ferguson, "Predictive policing systems offer a way seemingly to turn the page on past abuses, while still legitimizing existing practices" [160].

6.2 Responding to Predictions

More fundamental than biases within data are the politics embedded within the algorithms. For although designing algorithms appears to be a technical task, the choices made can have vast social and political impacts. All too often, algorithms that promise efficiency as a neutral good reflect the priorities of existing institutions and power structures. In privileging police efficiency in reducing crime rates over alternative goals such as improving neighborhood welfare with social services, supposedly neutral models further entrench the role of police as the appropriate response to social disorder. In that sense, predictive policing is likely to have discriminatory impacts not just because the algorithms may themselves be biased but also because they are deployed to grease the wheels of an already discriminatory system.

Rather than rush to adopt machine learning, we must ask: What goals should we pursue with the aid of predictive algorithms? How should we act in response to the predictions that are generated? How can we alter social and political conditions so that the problem we want to predict simply occurs at lower rates? Not every application of machine learning is inevitably biased or malicious, but achieving benefits from machine learning requires that we debate—in political rather than technical terms—how to design algorithms and what they should be deployed to accomplish.

Policing is not the only or the most effective way to curb crime and aid communities—in fact, as the police scholar David Bayley explains, "one of the best kept secrets of modern life" is that "[t]he police do not prevent crime" [32]. For example, a 2017 study found that proactive policing "may inadvertently contribute to serious criminal activity" and "curtailing proactive policing can reduce major crime," suggesting that one of the most common (and discriminatory) police practices does not even achieve its stated purpose of reducing crime [470].

Although police possess means and powers to deter and punish certain criminal activity, they are ill-equipped to take on the full range of issues with which they are increasingly required to deal: poverty, homelessness, mental health and drug crises, isolated neighborhoods with poor education and limited job opportunities. These

issues would be better addressed by alternative interventions. It is only by starting with a comprehensive and compassionate understanding of what factors lead to contact with the criminal justice system and what tactics can be used in response that algorithms can truly help generate a more just city.

Instead of conceiving more holistic approaches that capture the complexity of the world, however, predictive policing efforts tend to adopt visions of society that fit the simple presumptions within models and to presume that the only possible social change is to make policing more efficient by using data and algorithms. A person's or neighborhood's "risk" of crime are treated as inherent attributes of those people or places.

For instance, one statistician describes the task of person-based risk assessments in vivid terms: "We have Darth Vaders and Luke Skywalkers, but we don't know which is which" [21]. The goal is to distinguish Vaders from Skywalkers. Although this description helps explain how the algorithm works, it also oversimplifies the social complexities of characterizing a person's "risk." The world cannot be broken down into people who want to destroy the universe and those who risk their lives to save it. Unwittingly, this analogy highlights the fallacies of such simplistic thinking. As one critic writes, "Darth Vader wasn't an unimpeachably evil individual. At one point he was an innocent little boy who grew up in some dire circumstances" [205]. Rather than question why people make certain decisions or end up in particular situations—and attempt to push them toward positive outcomes—this approach to developing risk assessments presumes that people are fundamentally either good or bad, and that our task is simply to determine whom to punish. All we can do is follow the binary representations defined by the algorithm.

An ambitious effort along these lines is to predict whether newborn babies in Norway will commit a crime before turning eighteen, from information such as where that baby lives and who its parents are [58]. If the same approach were taken in the United States, there is little doubt that a machine learning algorithm could distinguish with reasonable accuracy between people who will be arrested and those that will not. After all, a government report estimated that of male babies born in 2001, one of out every three blacks, compared to only one out of every seventeen whites, would go to prison at some point during his life [43]. Given those stark statistics, we

don't need cutting-edge algorithms to predict who will be arrested.

Just because we can predict a certain outcome does not mean we should consider that outcome to be inevitable or just. That a model could predict a baby's future criminality reflects the vast inequalities of justice and opportunity in society, not the inherent nature of certain people. In just the last century, African Americans have, among many injustices, been excluded from government programs that provided loans for education and housing and been funneled into prisons through the war on drugs [10, 424]. The vast disparities in education, income, and crime that have resulted from these actions are not inevitable but the product of discrimination. A person or neighborhood's "risk," in other words, reflects the social and political conditions that shape behavior and outcomes.

A 2012 advertisement for IBM's Domain Awareness System portrays a similar perspective. The commercial follows two white men—the proverbial cop and robber—driving through city streets at night. The police officer provides a voiceover that begins as follows: "I used to think my job was all about arrests. Chasing bad guys. Now I see my work differently. We analyze crime data, spot patterns, and figure out where to send patrols." Relying on the advice of a computer in his police car, the officer reaches a convenience store just in time to thwart the would-be thief [242].

Although it tells an appealing story, IBM's ad demonstrates how predictive policing software both relies on and perpetuates simplistic notions of policing and crime. The officer's first two statements set up the rules of society: there are "bad guys" who commit crime and police (the implied "good guys") whose job it is to arrest them. This story presents another Luke Skywalker versus Darth Vader scene, with no backstory (for apparently none is needed) to explain how each person came to their present roles. In this way, in addition to completely exaggerating what algorithms are capable of—no system can predict crime at scale with anywhere near the level of precision depicted—IBM's ad ignores all of the social and political dynamics that underlie crime and policing. The society portrayed in this vignette has no poverty, no segregation, no stop-and-frisk—in fact, because every character is white, it has no racial dynamics at all. We are left with a facile and pernicious conclusion: because

of the presence of "bad guys," crime is an inevitable phenomenon that can be prevented only by police who possess the necessary information.

Predictive policing thus suffers from a gaping divide between the problem being solved and the problem that needs solving. Owing to their focus on technology, many believe that the issues of policing stem from poor information about when and where crime will occur in the future. This is a problem that (at least in principle) new technology can solve. But as Alex Vitale argues in *The End of Policing*, "The problem is not police training, police diversity, or police methods. …The problem is policing itself." Tracing the history of policing from its roots to the present day, Vitale concludes: "American police function, despite whatever good intentions they have, as a tool for managing deeply entrenched inequalities in a way that systematically produces injustices for the poor, socially marginal, and nonwhite" [501].

In the hands of police, even algorithms intended for unbiased and nonpunitive purposes are likely to be warped or abused. In Chicago, for example, an algorithm conceived to reduce violence was perverted—through police control—into a tool for surveillance and criminalization. Drawing on research regarding how gun violence clusters in social networks (Papachristos and Wildeman, 2014), the Chicago Police Department (CPD) developed an algorithm to identify the people most likely to be involved in gun violence. And although the original stated intention for this "Strategic Subjects List" (SSL) was to prevent violence, it has largely been used as a surveillance tool that many believe disproportionately targets people of color [188]. A RAND evaluation concluded that the SSL "does not appear to have been successful in reducing gun violence"; instead, "the individuals on the SSL were considered to be 'persons of interest' to the CPD" and were more likely to be arrested [434].

Following the analysis of this chapter, the next chapter articulates a novel and alternative approach to predicting and responding to violence in Chicago. The methods described in the next chapter are grounded in two central insights. First, that risk is a product of social and structural—rather than natural and individual—conditions. Second, that predictions of risk must be responded to with rehabilitative and supportive social services rather than with punitive policing.

Chapter 7

Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence

7.1 Introduction

In 2013, 11,208 people in the United States were murdered with a firearm and approximately 62,220 others were injured in non-fatal gun assaults [167]. Although mass shootings are often the focus of public attention, the vast majority of gun murders and assaults occur in everyday incidents involving a small number of people (typically two) [513]. Furthermore, gun violence tends to concentrate within socially and economically disadvantaged minority urban communities where rates of gunshot injuries far exceed the national average [348, 390] and where young Black men experience rates of gun homicide ten times higher than their white counterparts [513].

The media, politicians, and academics alike often describe gun violence in the U.S. as an "epidemic" [51, 85, 95, 453, 513, 524], implying concern over its alarmingly high levels as well as the possibility of its spread. Although gun violence's stubborn persistence in certain communities might be more accurately described as an endemic

[85], the public emphasis on epidemics has inspired research on the mechanisms through which violence might be transmitted [345, 385, 524]. The most common approach measures the spatial diffusion of gun violence from neighborhood to neighborhood [91, 345, 348, 524]. Although this spatial approach often discusses interpersonal relationships related to gang activity [478, 524] or drug markets [90] as the drivers behind the diffusion of gun violence, the statistical models themselves presume that violence might be conceptualized an airborne pathogen (such as influenza) moving between neighborhoods, and that can be "caught" by inhabiting locations with high incidence rates.

Recent thinking suggests, however, that many of the processes that we attribute to geography might occur in part due to the interpersonal ties underlying social networks [429]. Research on gun violence in Chicago, Boston, and Newark has found that gunshot victims are highly concentrated within networks, along with cross-sectional evidence that such concentration is related to social contagion, i.e., the spreading of beliefs, attitudes, and behaviors through social interactions [381, 385, 382, 480]. Furthermore, social networks are fundamental in diffusion processes related to diverse areas such as behaviors [75], opinions [24, 44], HIV [4], obesity [83], and depression [423]. Taken together, these studies suggest that the diffusion of gun violence might occur through person-to-person interactions, in a process akin to the epidemiological transmission of a bloodborne pathogen (such as HIV). Contagion via social ties, then, may be a critical mechanism in explaining why neighborhoods matter when modeling the diffusion of crime and, perhaps more importantly, why certain individuals become victims of gun violence while others exposed to the same high-risk locations and situations do not.

To study the role of social influence in gun violence, we examined a particular interaction between individuals: being arrested together for the same offense, a behavior known as co-offending. Co-offending typically occurs between people who share strong pre-existing social ties [502] and is driven by social processes that amplify risky behaviors (criminal or delinquent acts that might lead to arrest, including violent victimization and offending) [222, 223, 481, 502, 521]. Like other social behaviors such as needle-sharing [282] and sex [4, 33], co-offending may reveal patterns of social interactions that influence how victimization spreads [159, 223, 374, 472, 502]. We

hypothesized that a person becomes exposed to gun violence through social interactions with previous victims: someone who has been shot may be more likely to be embedded in the networks and environments in which guns are present and gun violence is likely to erupt. Associating with victims of gun violence, and specifically co-engaging in risky behaviors with them, therefore may expose individuals to these same behaviors, situations, and people that in turn increase the probability of victimization.

Our study directly assessed the efficacy of treating the diffusion of gunshot victimization as an epidemiological process that spreads through social networks. Our central hypothesis was that when someone in your network becomes a victim of gun violence, your risk of victimization temporarily increases. We hypothesized that predictive models incorporating social contagion would outperform models considering only individual and ecological risk factors in predicting future gunshot victims. Modeling the precise social dynamics of victimization could represent an important advance in treating gun violence as a public health epidemic. By uncovering high-risk individuals and transmission pathways that might not be detected by other means, a contagion-based approach could detect strategic points of intervention that would enable measures to proactively reduce the trauma associated with gun violence rather than react to past incidents. Importantly, such a contagion-based approach is *victim-centered*, and as such has the potential to move the larger public dialogue on gun violence away from efforts that rest largely on geographic or group-based policing efforts that tend to disproportionately affect disadvantaged minority communities.

We tested our hypothesis in Chicago, IL, a city whose well-documented patterns of gun violence are emblematic of the epidemic described above and whose rates of gun violence are more than three times the national average (Figures 7.1 and 7.1) [42, 206, 347, 384, 429, 371]. Although Chicago does not have the highest urban per capita homicide rate, the city has a long history of violence and consistently tallies a greater number of homicides than any other city in the United States [124].

As in other major U.S. cities, violent gun crime in Chicago is intensely concentrated in a small number of socially and economically disadvantaged neighborhoods (where homicide rates can be upwards of 75 per 100,000

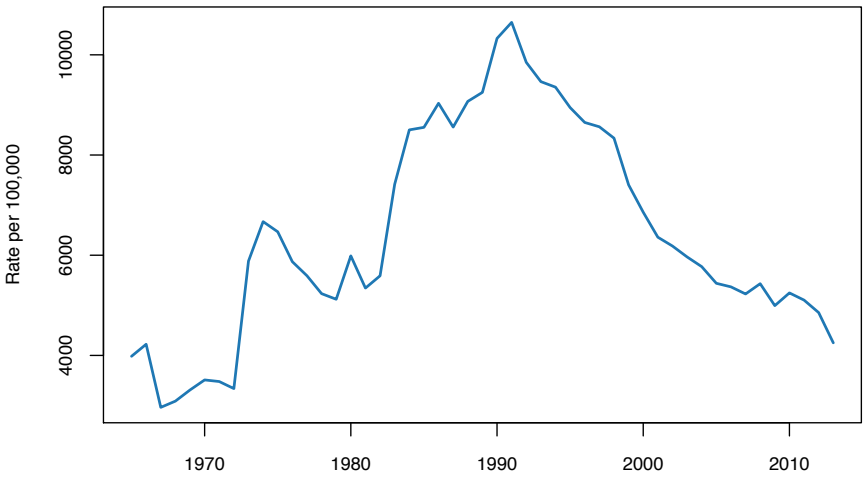

Figure 7.1: Index crime in Chicago (rate per 100,000), 1965 to 2013. Index crimes include all murders, criminal sexual assaults, aggravated assaults/batteries, burglaries, thefts, robberies, arsons, and motor vehicle thefts. Crime rose throughout the 1970s and 1980s, peaking in 1991 with a rate of 10,647.9 per 100,000 people. Crime in Chicago has since declined steadily, with a rate of 4251.2 per 100,000 in 2013. Data come from the FBI Unified Crime Reports [371].

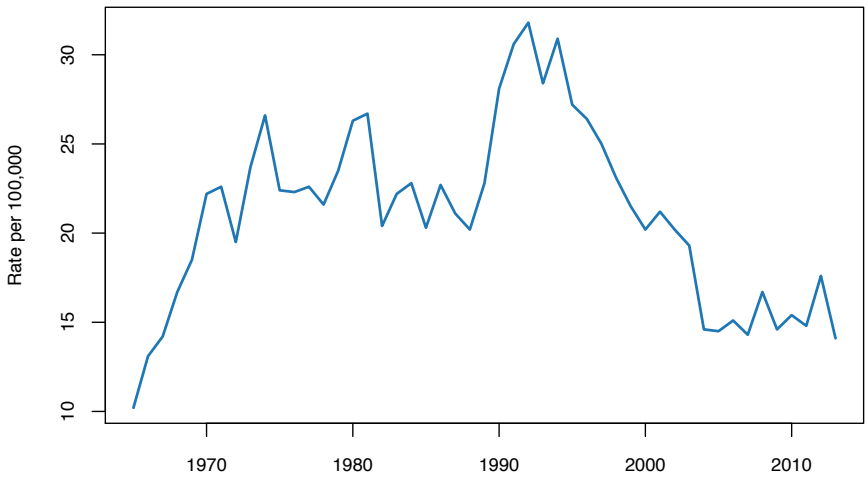

Figure 7.2: Homicide in Chicago (rate per 100,000), 1965 to 2013. Homicide rates between 1965 and 2013 follow a similar pattern as index crime rates, peaking in the early 1990s (with a rate of 31.8 per 100,000 in 1992) and declining steadily since then. The homicide rate in 2013 was 14.1 per 100,000 people, the lowest since 1966. Homicide data from 1965 to 1994 were provided by Carolyn Rebecca Block and Richard L. Block through the National Archive of Criminal Justice Data [42]. Detailed data on homicides from 1995 to 2010 were provided by the Chicago Police Department.

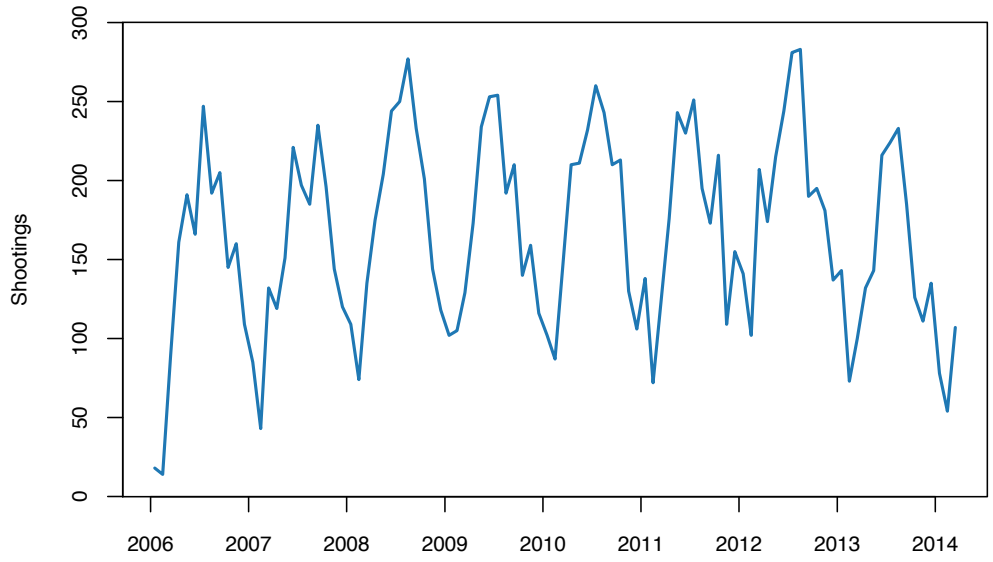

Figure 7.3: Monthly counts of fatal and non-fatal gunshot injuries in Chicago, 2006 to 2014. The number of shootings per month varies widely depending on the time of year: violence peaks in the summer and declines in the winter. In 2008, for example, the number of shootings per month varied from 74 in February to 277 in August.

people) [42, 384, 438]. Furthermore, gun violence victimization is concentrated in a small social networks: a recent of non-fatal gunshot victimization in Chicago from 2006 to 2014 found that greater than 70% of all victims could be located in networks containing less than 5% of the city's population [385]. The current study examines the extent to which individual gunshot victimization in Chicago might be explained as a process of epidemiological transmission between individuals in these networks.

The main variable of interest in the present study is fatal and non-fatal gunshot victimization, excluding self-inflicted and accidental gunshot injuries as well as legal interventions (i.e. police-related shootings). Figure 7.3 plots the monthly combined number of fatal and non-fatal gunshot injuries during the observation period, 2006 to 2014. The expected seasonal variation of gun violence [328, 329], with peaks in the summer months, is also apparent. Average monthly rates during the study period ranged from 71.25 shootings in February to 245.5 in July.

7.2 Methods

7.2.1 Data

We examined all recorded fatal and non-fatal gunshot injuries in Chicago from 2006 to 2014 among the population of individuals arrested during this time period. Data come from two different sources provided by the Chicago Police Department through a nondisclosure agreement (and approved by the Yale Institutional Review Board):

1. All 1,189,225 arrests recorded by the police from January 1, 2006 to March 31, 2014, involving 462,516 people (for comparison, the adult population of Chicago totals approximately 2.1 million). Arrest data are recorded at the incident level and contain social and demographic information on each reported individual including birthdate, race, ethnicity, sex, and gang membership (as identified by the police).

2. Detailed records for all 16,399 gunshot victimizations recorded by the police during the same time period, excluding suicides, accidents, and shootings that occurred during legal interventions (i.e., shootings involving law-enforcement personnel). These records consist of 13,877 non-fatal and 2,522 fatal shootings, affecting 14,695 people; 1,498 people were shot on more than one occasion. Among all shooting victims, 90.2% were arrested during the study period and could be located in the arrest data.

Events and individuals are uniquely identified across both datasets using internal alphanumeric codes created by CPD (which we refer to as Event Codes and Identity Codes, respectively), thereby allowing us to match events and people over time and across datasets.

These data have important limitations. First, police data have known biases, including: (a) undercounts of the true volume of crime because most crimes go unreported; (b) problems caused by data-entry errors or the use of aliases; and (c) biases in criminal justice practices and polices, including racial and neighborhood profiling, that might skew the true geographic and socio-demographic distribution of crime [36, 273, 477]. Regarding this last

point, we make no claims of whether arrests were justified, but simply rely on them as the systematic recording of an observed behavior. Second, since crime is generally underreported, our co-offending data most likely underestimates the social ties related to risky behavior. And, third, without comparative data from other cities, it is difficult to know how representative the Chicago co-offending network is of co-offending more generally.

7.2.2 Co-Offending Network

Figure 7.4 illustrates how the co-offending network was created. We generated a social network from the data by identifying all unique individuals arrested during the study period and connecting them via "edges;" that is, a relationship between pairs of individuals defined by being arrested together for the same offense (a behavior known as co-offending) at least once during the study period (Section 7.2.1). This network contained 462,516 individuals, 467,506 edges, and 13,252 victims.

Forming the Network

We used the arrest records to generate a bipartite network that connects arrest events and people (Figure 7.4). That is, the network connects each person to every arrest in which he or she was involved. Equivalently, the network connects each arrest event to all of the people arrested. The network is clearly bipartite since people cannot be linked with people and arrests cannot be linked with arrests. This network has a total 1,189,225 arrest event nodes, 462,516 person nodes, and 1,458,957 edges.

We performed a bipartite projection on the person nodes to obtain a one-mode social network, where nodes represent each person who was arrested during the study period. This network contained 462,516 nodes and 467,506 edges. Unweighted edges connect every pair of people who were arrested in the same event during the study period, connecting individuals based on their association with the same crime.

Edges connected pairs of individuals who co-offended together at some point during the study period. Due to the one-mode structure, incidents in which more than two people co-offended together were represented by

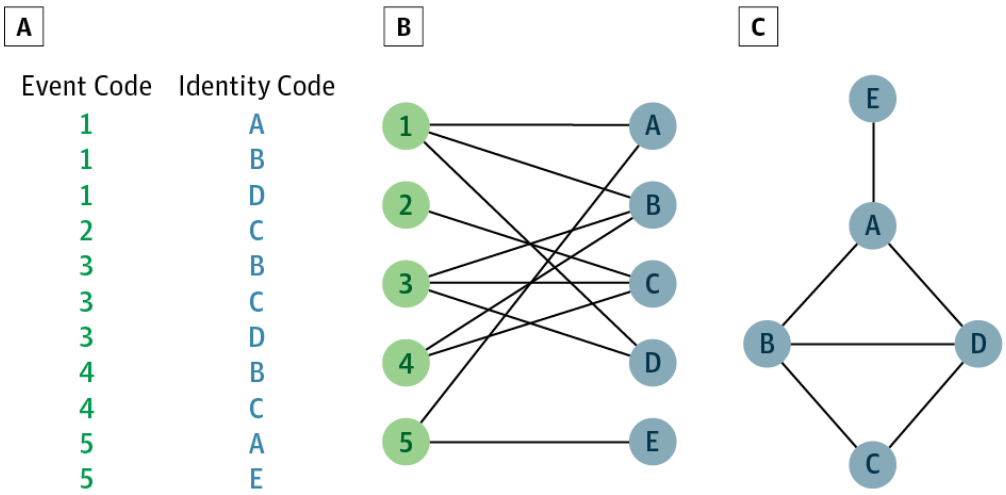

Figure 7.4: Co-offending Network Generation Process. (A) Example of raw data and its structure, in which Event Codes (ECs) mark specific arrest events and Identity Codes (ICs) represent unique individuals. Each entry represents a single individual arrested in a specific incident. (B) Bipartite (i.e., two-mode) network between offenses (green) and people (blue), generated by using the data from (A) as an edge list (in which each row represents a pair of nodes that are connected by an edge). (C) Person-to-person (i.e., one-mode) co-offending network, generated by performing a bipartite projection on the network from (B). Nodes represent unique offenders and edges connect offenders who were arrested for the same incident. Note that the network shown in this panel is un-weighted, meaning that every edge has identical weight=1, even for pairs of individuals who were arrested together multiple times.

edges between every pair of individuals involved rather than all individuals to a common incident. More than two-thirds of the arrests involving multiple people had only two participants, however. Another 18% contained three people, leaving only 13% of arrests that involved more than three people. This shows that our one-mode co-offending network is a reasonable representation of co-offending dynamics.

We treated the co-offending network as static rather than forming each edge at the date of first co-offense. Although the co-offending events occur at specific points in time, previous research on co-offending has shown that the individuals involved typically already have close existing relationships [502]. Because we can identify the presence of these relationships but not the date when those relationships formed, we generate a static network that includes every co-offense throughout the study period.

While it is possible to build a weighted social network with edge weights corresponding to the number of co-arrests between individuals, we did not do this for three reasons. First, it is difficult to determine if two people appear together in the arrest data multiple times because they actually co-offended together multiple

times or simply faced multiple charges from the same co-offense. Second, there are very few edges between individuals who co-offended multiple times together—a finding consistent with prior research on co-offending networks [330]. 94% of edges have weight=1, 5% of edges have weight=2, and all larger edge values account for the remaining 1% of edges. Finally, we found no evidence that high-weight edge facilitate the transmission of gunshot victimization. For each edge weight represented in the network, we looked at the percentage of pairs where both individuals were infected. The probability that such a pair exists actually goes down as the edge weight increases. In particular, out of the 269 highest-weight edges (weight>5), there is not a single pair where both people were infected. This leads us to believe that the few high-weight edges that exist have little or no special effect on the contagion of violence.

Adding Victim Attributes

We used gunshot victimization records to determine our dependent variable of whether or not any individual in the data was a victim of a fatal or non-fatal gunshot injury during the study period. For each victim, we record the date of every fatal and non-fatal gunshot victimization associated with that individual. Eleven percent (N=1,251) of the victims of non-fatal gunshots had multiple victimizations during the study period, with a maximum of five. Twelve percent (N=247) of the victims of fatal gunshots had previously during the study period been the victim of a non-fatal gunshot.

Although we restricted ourselves to gunshot victims who are also in the co-offending network, we still captured the vast majority of victims in our analysis: 93% of nonfatal victims members of the co-offending arrest network, and 80% of fatal victims are in the network.

The Largest Connected Component

Decomposing the co-offending network into disjoint connected components yielded many small components and one giant component. This is similar to the pattern observed in other empirical networks [3, 7]. More than

half of the nodes (56%) are isolated, corresponding to people who were never arrested with anyone else. Of the 284,876 connected components, only one contains more than 30 nodes. This largest connected component contains 30% of the nodes in the network (N=138,163), 89% of the edges (N=417,635), and 74% of the victims (N=9,773). As is standard in social network analyses [3], we take the largest connected component (LCC) as the focus of our study.

The largest connected component resembled a typical social network. The network's degree distribution followed a power-law distribution with scaling exponent 1.39. This means that the LCC is a scale-free network, which is very common among social networks [7]. The LCC has a clustering coefficient of 0.6 and an average path length of 8.3. In comparison, an Erdös-Réyni random graph of the same size has a clustering coefficient of 0.00003 and an average path length of 6.78. Since the LCC is highly clustered with a similar average path length compared to a random graph, it is a small-world network [504].

We restricted our analysis to the network's largest connected component, which contained 29.9% of all the arrested individuals (N=138,163) and 89.3% of all the co-offending edges (N=417,635). Consistent with previous research on the concentration of gun violence within co-offending networks [385], the largest connected component contained 74.5% of gunshot victimizations of arrested individuals (N=11,123 victimizations, affecting 9,773 people). We henceforth refer to this component as the Network.

7.2.3 Homophily, Confounding, and Contagion

Understanding how gunshot victimization might make its way through a network requires understanding different reasons for how patterns of gunshot violence might emerge in a network: failing to account for all possible explanations can lead to overestimating the effects of social contagion [18, 93, 448]. We consider three potential explanations: individuals associate with similar peers (homophily), individuals are exposed to the same environmental factors (confounding), and individuals influence one another's behavior over time (contagion) [14, 18, 83, 93]. To distinguish between these explanations, we analyzed the temporal patterns of victimization

with those generated by simulations that account for homophily and confounding but not contagion. We ran 10,000 Monte Carlo simulations of the study period, assigning to each victim a new victimization date that is consistent with his or her exposure to violence based on risk factors and environmental influences. By shuffling the infection dates between victims as described below, the simulations generated a set of networks that 1) retained the aggregate patterns of gun violence, as measured by the number of victimizations each day, and 2) accounted for the effects of homophily and confounding but assume no social contagion.

Homophily would explain the temporal clustering of victims in the network if people co-offend with others who have similar risk factors and therefore are likely to be shot at similar times. Many prior studies have shown strong relationships between certain risk factors and exposure to violence [480], a relationship that our data corroborates. In our simulations we controlled for whatever traits cause two individuals to co-offend together by holding constant the network structure and victim identities: each victim has the same neighbors in both the real and simulated data. Confounding would explain the pattern of victimizations if features such as age and neighborhood expose similar individuals to violence at the same time. Our simulations controlled for confounding by shuffling victimization dates only between individuals who are the same age, gender, and ethnicity; live in the same neighborhood; and both either belong or do not belong to a gang (if an individual does not match with anyone else across all of these features, that person's infection date is not altered). We also controlled for the fact that violence rates fluctuate, following a predictable seasonal trend of rising in the summer and declining in the winter [328, 329]. Furthermore, some years have more incidents of violence than others and crime in the US and Chicago declined during the observation period [380, 527]. In order to control for violence rates over time, we simulated the exact same number of infections per day as observed in the data. Together, these controls ensured that we accurately represented each person's exposure to violence as it relates to individual and environmental risk factors.

This approach allowed us to determine the extent to which the observed concentration of victims could be explained without any social contagion. If the concentration of victims was primarily due to homophily and

environmental confounding, the simulations would accurately recreate the observed pattern of gunshot violence. On the other hand, if social contagion was responsible for some victimizations, we expected to see that the observed victimizations appear closer together in time than the simulations could explain.

Because we held constant the set of victims and infection dates, we could simulate an infection process that lacks social contagion by shuffling the matching between victim identity and victimization date. Given our method's similarity to the previously-developed "Shuffle Test" [14], we refer to our approach by the same name.

Our Shuffle Test ran as follows:

1. Take the LCC and identify the gunshot victims from the data.

2. Divide all the victims into groups that share the same birth year, gender, ethnicity, residential neighborhood, and gang membership status (i.e. belong to a gang or not).

3. Within each group, randomly permute the infection dates associated with each individual. Individuals in groups by themselves retain the same infection date.

This yielded a new version of the LCC, with the same victim population and overall set of infection dates as the raw data. Each victim was infected at different times during the simulated study period compared to the observed data, but in a manner that is consistent with the rate at which he or she was exposed to violence.

For each simulation, we measured how many days passed between infections within every pair of first-degree associates who were both victims (N=9,568). If one or both victims were infected multiple times during the study period, we take the minimum time difference between infections. As our network and set of victims are fixed based on the data, the quantity and identities of such pairs remain constant in every simulation. If these pairs of victim are shot equally close together in time in the data and simulations, then we will be able to conclude that homophily and confounding are sufficient to explain the data. Alternatively, if the data exhibits a higher degree of temporal clustering, this will imply that explanations beyond homophily and confounding are necessary.

7.2.4 Social Contagion Model

We modeled the contagion of violence using a multi-dimensional Bayesian Hawkes process over the co-offending network. We first present the general definition of Hawkes Processes, then instantiate and adapt it to the contagion of gun violence over a network.

Hawkes Processes

Hawkes processes are a class of self-exciting temporal point processes originally introduced by Alan Hawkes in the early 1970s [220], and have recently become common as a way to model contagion and diffusion processes. Applications include the spread of seismic events [322], information spread in social networks [157], and stock market trading dynamics [304].

A convenient way to describe temporal point processes is through their conditional intensity function, which describes the instantaneous probability of occurrence of an event at any given time t. For Hawkes processes, the conditional intensity function can be written as the sum of endogenous time-varying intensities (capturing the intra-network influence of the events preceding time t) and an exogenous intensity (capturing the influence of all extra-network factors).

Formally, for a D-dimensional Hawkes process with N infection events, let us introduce the set of events $\varepsilon = (t_i, k_i)_{1 \leq i \leq N}$ where t_i denotes the time of event i and k_i the dimension (or coordinate) on which it occurs. The conditional intensity function is defined as follows:

$$\lambda_k(t) = \mu_k + \sum_{i=1}^{N} \varphi_{k_i,k}(t - t_i) \tag{7.1}$$

where $M = (\mu_k)_{1 \leq k \leq D}$ is the vector of exogenous intensities (also known as background rates) and the functions $\Phi = (\varphi_{i,j})_{1 \leq i,j \leq D}$ is the matrix of endogenous kernel functions (also known as exciting functions). For a pair of coordinates (u, v), $\varphi_{u,v}(t)$ models the influence of coordinate u over coordinate v after t time has passed since

u was infected. The kernel functions are non-negative and causal (i.e. $\varphi_{u,v}(t) = 0 \forall t < 0$). In particular, this implies that the summation in Equation 7.1 is only over the indices i such that $t_i < t$.

From this definition, we see that the Poisson process can be characterized as a special case of the Hawkes process, with a constant exogenous intensity and no dependence on past events. That is, $\lambda(t) = \lambda$.

I refer the reader to other sources [114, 412] for a formal discussion of the conditional intensity function and its proper interpretation in a Hawkes process. From these we apply the following formula for the log-likelihood of events $\varepsilon = (t_i, k_i)_{1 \leq i \leq N}$ given M and Φ over observation period $[0, T]$:

$$\mathcal{L}(\varepsilon | M, \Phi) = \sum_{i=1}^{N} \log \lambda_{k_i}(t_i) - \sum_{k=1}^{D} \int_0^T \lambda_k dt \tag{7.2}$$

The first sum calculates the log-likelihood of every infection event that *did* occur, and the second sum calculates the log-likelihood that each individual was *not* infected at all other times.

Contagion of Gun Violence as a Hawkes Process

We model the contagion of gun violence as a Bayesian Hawkes Process by defining the following features: each network vertex (i.e. each individual) occupies its own coordinate of the Hawkes Process and each gunshot victimization is an event of the process occurring on the coordinate that corresponds to the victim (repeated victimizations of the same individual correspond to multiple events on the same node, and are treated the same as single victimizations).

Exogenous Intensity. We assume that the exogenous intensity is the same for every individual in the network, and attribute the observed fluctuations of violence rates (Figure 7.3) to a seasonal effect independent of peer contagion. For this reason, we fit a time-varying function $\mu(t)$ to the data and use it for the common exogenous intensity.

Endogenous Exciting Functions. The exciting function $\varphi_{u,v}(t)$ models the effect of person u on person v after t time has passed since u was infected and captures two common assumptions regarding the spread of

contagions (Figure 7.5).

First is time. Consistent with previous models used to infer the spread of contagions over social networks [157, 184, 303], we assume that the impact of earlier infections on future events decays as the time passed since the original infection increases. Additionally, influence can only travel forward in time: an infection has no impact on those that came before it. As is common for Hawkes processes [114, 157, 184, 303, 304, 412], we assume an exponential decay and obtain the following formula for the temporal component of the exciting functions:

$$f_\beta(t) = \beta e^{-\beta t} \text{ if } t > 0, 0 \text{ otherwise} \tag{7.3}$$

Second is network structure. Epidemiologists commonly assume that contagious events are localized and that the transmission probability increases closer to the source [203, 447, 467, 496]. In our case, we assume that violence is more likely to spread between people who are closely linked in the network and measure the distance between individuals based on network topology. Based on previous studies of violence in social networks [83, 385], we assume that infections are able to be transmitted across a network distance of up to three degrees of separation; people who are further away in the network have no effect on one another. Hence, we obtain the following formula for the structural component:

$$g_a(u, v) = a\,\text{dist}(u, v)^{-2} \text{ if } \text{dist}(u, v) \leq 3, 0 \text{ otherwise} \tag{7.4}$$

Finally, we obtain the exciting function by combining the above two components:

$$\varphi_{u,v}(t) = f_\beta(t) g_a(u, v) \tag{7.5}$$

A

	Identity Code (IC)	Victim	Infection date
1	A	TRUE	2
2	B	FALSE	NA
3	C	FALSE	NA
4	D	TRUE	4
5	E	FALSE	NA

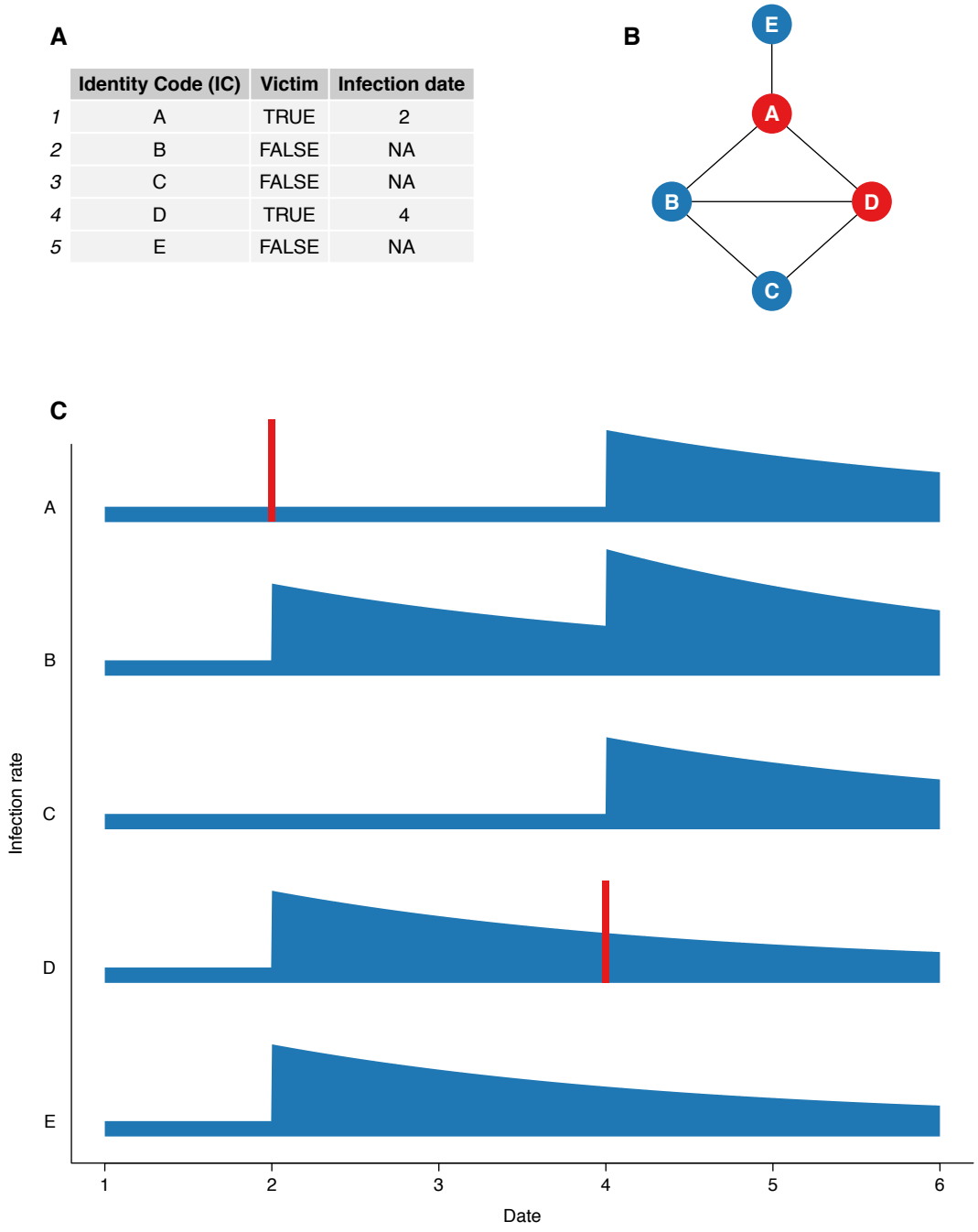

Figure 7.5: Hawkes model dynamics over an example network. (A) A table of identities and whether each individual was a gunshot victim. The infection time for each victim is also recorded. (B) The co-offending network of individuals in (A), with victims marked in red and non-victims in blue. (C) The infection rates of each person in the network over a five-day period. Each individual is initially susceptible to infection only due to a small background rate, based on exogenous features, that is constant across the population. When individual A is infected on day 2 (marked with a red line), it causes a spike in the infection rate of its three neighbors: B, D, and E. The impact of this infection decays over time. Because a node cannot generate further infections in itself, A's infection rate does not change when it is infected. Node D is infected on day 4 (marked with another red line), causing the infection rates of A, B, and C to spike. Because the effects of peer contagion are additive and B is connected to both infected nodes, B has the highest infection rate after D is infected.

Model Likelihood

Using Equation 7.2 and the model presented in Section 7.2.4, we can now write the log-likelihood of observed infection events $\varepsilon = (t_i, k_i)_{1 \leq i \leq N}$. V denotes the set of vertices in the network, and $[0, T]$ marks the study period.

Since some individuals were the victims of fatal gunshots during the study period, they were not susceptible to infection during the entire study period. For these victims, the second summand of Equation 7.2 only needs to be integrated until their time of death. Denoting by T_v the time of death of vertex v ($T_v = T$ if the individual didn't die during the study period), we obtain:

$$\mathcal{L}(a, \beta, \mu | E) = \sum_{i=1}^{N} \log \lambda_{k_i}(t_i) - \sum_{v \in V} \int_0^{T_v} \lambda_v(t) dt \tag{7.6}$$

7.2.5 Inferring Model Parameters

In this section, we describe how we learned the optimal parameters to describe the Hawkes model.

Exogenous Intensity

Because the seasonal variations in gunshot rates remain consistent throughout the study period (Figure 7.3), we assume these are not purely driven by noise or social contagion. We model these seasonal variations with a periodic sinusoidal function.

Let $M(t)$ denote the expected number of total victimizations occurring on day t. We assume the following form:

$$M(t) = A\left[1 + \varrho \sin(\omega t + \varphi)\right] \tag{7.7}$$

Because violence fluctuates annually, we know that the period is one year, i.e. $\omega = 2\pi/365.24$. We learn the remaining three parameters $\{A, \varrho, \varphi\}$ using non-linear least squares estimates with the Gauss-Newton algorithm.

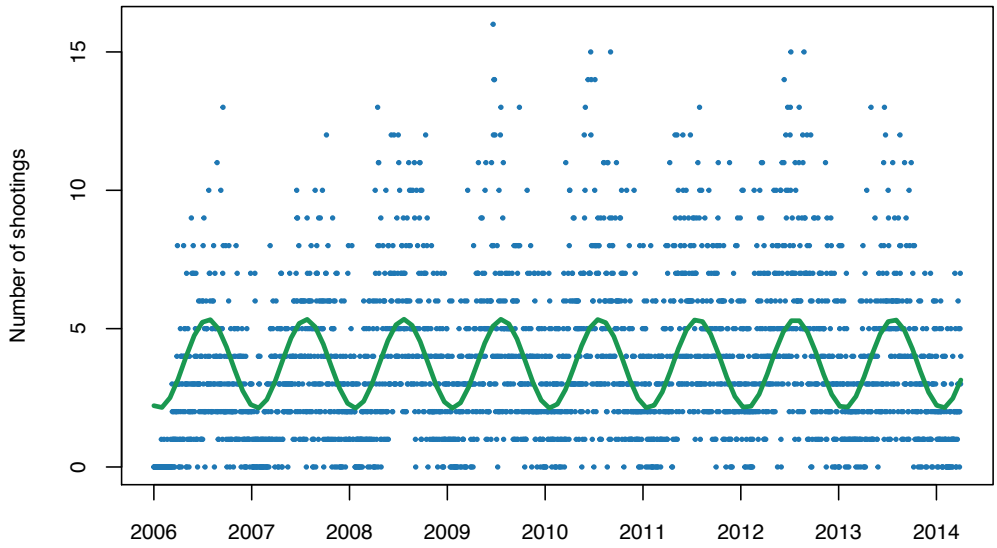

Figure 7.6: Shootings per day and best-fit curve during the study period. Each blue dot represents the number of shootings (fatal and non-fatal) on a single day. Values ranged from 0 (N=280, 9.3%) to 16 (N=1), with a mean of 3.7 and median of 3. In order to model how violence rates vary over time, we fit a sinusoidal curve to this data (in green).

This yields:

$$M(t) = 3.73\left[1 + 0.43\sin\left(\frac{2\pi}{365.24}t + 4.36\right)\right] \tag{7.8}$$

Figure 7.6 depicts the number of infections on each day of the study period along with the function $M(t)$.

Because we do not yet know the importance of the exogenous intensity in spreading gunshot violence, we keep only $\{\varrho, \varphi\}$ from the fitted parameters. In other words, we only keep the parameters characterizing the seasonal fluctuations; the base amplitude A of the exogenous intensity will be inferred together with the kernel function parameters in the following section.

Finally, we relate the aggregate number of infections $M(t)$ to the node-level exogenous intensity $\mu(t)$. By definition:

$$M(t) = \sum_{v \in V} \int_{t-1}^{t} \mu(s)ds = |V| \int_{t-1}^{t} \mu(s)ds \tag{7.9}$$

where we used that the exogenous intensity is identical across all nodes. Assuming that $\mu(t)$ is approximately constant over the course of one day, we get $M(t) = |V|\mu(t)$. Hence we obtain the following form for the

exogenous intensity:

$$\mu(t) = \mu_0 \left[1 + 0.43\sin\left(\frac{2\pi}{365.24}t + 4.36\right)\right] \tag{7.10}$$

where $\mu_0 = A/|V|$.

Learning the Optimal Model Parameters

Using the exogenous intensity, the log-likelihood now depends on three parameters $\{a, \beta, \mu_0\}$. Finding the maximum likelihood estimate of these parameters amounts to solving the following optimization problem:

$$\hat{a}, \hat{\beta}, \hat{\mu} = \underset{a,\beta,\mu}{\arg\max}\ \mathcal{L}(a, \beta, \mu|\varepsilon) \tag{7.11}$$

Unfortunately, the function $\mathcal{L}(a, \beta, \mu|\varepsilon)$ is not jointly concave in its three arguments. We will, however, exploit the following fact.

Proposition 1. The function $(a, \mu_0) \mapsto \mathcal{L}(a, \beta, \mu|\varepsilon)$ is concave.

Proof. Expanding the terms in Equation 7.6, it is clear that the second sum is linear in $\{a, \mu_0\}$. Hence it is sufficient to show that for $1 \leq i \leq N$:

$$h(a, \mu_0) = \log\left(\mu_0\left[1 + 0.43\sin\left(\frac{2\pi}{365.24}\right)\right] + \sum_{j:t_j<t_i} a\text{dist}(u, v)^{-2} f_\beta(t)\right) \tag{7.12}$$

is concave. For this, we see that the operand of the log function is linear in $\{a, \mu_0\}$. By composition with the concave function log, we obtain that h is concave and thus conclude the proof. □

We observed numerically that has many local optima; hence we solve Equation 7.11 using the following heuristic:

1. We perform a brute force grid search to locate good starting points for the refining heuristic.

2. Starting from the best point obtained during the first step, we refine the solution by alternated minimiza-

tion. First, optimize over $\{a, \mu_0\}$ for a fixed value of β. By Proposition 1 we were able to use standard convex optimization (gradient descent, in this case) to solve this step exactly. Second, optimize over β for a fixed value of $\{a, \mu_0\}$, using simulated annealing.

Other heuristics were considered: using gradient descent as well for the optimization over β, or using global gradient descent to optimize over $\{a, \beta, \mu_0\}$ at the same time. All heuristics led to the same optimal solution, indicating that our initial grid search was precise enough to identify good starting points. We obtained the following values of the parameters at the optimum:

$$a = 7.82 \times 10^{-3}, \ \ \beta = 3.74 \times 10^{-3}, \ \ \mu_0 = 1.19 \times 10^{-5} \tag{7.13}$$

Validation on Simulated Data

In order to validate our approach for learning the Hawkes model parameters, we evaluated our method on synthetic data. Starting from the same co-offending network as in the dataset (i.e. the LCC), we generated synthetic contagion events by simulating the Hawkes contagion model described in Section 7.2.4 [356]. The model parameters are fixed to the values obtained in Equation 7.13.

We then computed the maximum likelihood estimator described in Section 7.2.5 on the synthetic contagion events and compared the inferred parameters to the true values used during the simulation. To analyze how our estimates converge as we observe more data, we truncated the synthetic dataset at increasing time horizons between 0 and 3,000 days (our study period spanned 3,012 days) and trained the maximum likelihood estimator separately on each truncated dataset.

We performed this procedure five times to generate five independent sets of synthetic contagion events (Figure 7.7). We observed that the inferred parameters for a and β vary for short study periods but quickly converged toward the optimal value as the study period increases. The learned parameters for μ_0 are close to optimal even for short study periods. After 3,000 days, all inferences for a were within 10.8% of the true value, all inferences

for β were within 12.7%, and all inferences for μ_0 were within 2.1%. The mean parameters for each parameter from the five trials at 3,000 days were all within 1.8% of the optimal value. These results indicate that our parameter inference method was able to reliably determine the parameters of a Hawkes model over the study period length used.

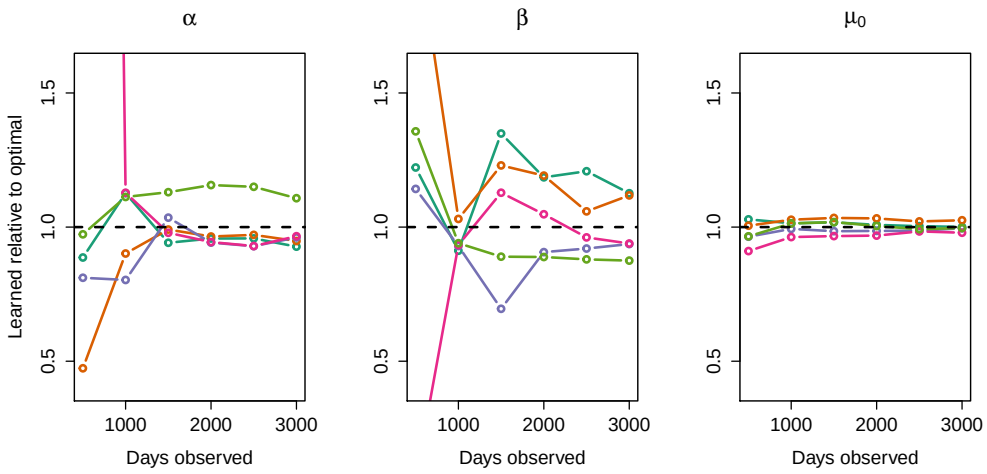

Figure 7.7: Learned parameters relative to optimal from five simulations of the Hawkes contagion process. We simulated five Hawkes contagion processes over the LCC using the parameters found in Equation 7.13. Using the method in Section 7.2.5, we learned the parameters that best describe the simulated data and compare these to the actual value. The dashed black lines indicate that the optimal result is for the learned parameters to be identical to the parameters actually used. Colored lines show the learned parameters relative to optimal for each simulated dataset as we observe a different number of days. We observe that the inferred parameters for a and β vary notably for short study periods but quickly converge toward the optimal value as the study period increases. The learned parameters for μ_0 are close to optimal even for short study periods. The means of the learned parameters from the five trials at 3,000 days are all within 2% of the optimal value, indicating that our parameter inference method is able to determine the parameters of a Hawkes model over the actual study period.

7.2.6 Inferring the Pattern of Infections

Using parameters calibrated on the observed data, our model calculated each person's exposure to gun violence based on the aggregate influence of social contagion and seasonal factors. For each gunshot victim who was influenced primarily by contagion, we identified which peer (the infector) was most responsible for causing them to become infected (i.e., shot). We then connected these infections from infector to victim to trace cascades of gunshot victimization through the network, i.e., chains in which one person becomes infected, exposing his or her

associates, who then may become infected and spread the infection to their associates, and so on (Section 7.2.4). It is important to note that the infector is not assumed to be the one who shoots the victim, but rather the one who exposes him or her to the risk of victimization.

Given fitted values of the parameters of the Hawkes contagion model, we then determined whether each infection event (t, v) was more likely to have been caused by the exogenous background rate or endogenous peer contagion. Using Equation 7.1, we compared the value of the exogenous and endogenous intensities at the time t of infection, and attributed the infection event to the larger of the two quantities. In other words, we compared:

$$\mu(t) = 1.19 \times 10^{-5} \left[1 + 0.43 \sin\left(\frac{2\pi}{365.24} t + 4.36\right)\right] \tag{7.14}$$

with

$$\sum_{u \neq v} \varphi_{u,v}(t) \tag{7.15}$$

and attributed the infection to peer contagion if Equation 7.15 >Equation 7.14.

For infection events (t, v) attributed to peer contagion, we could single out a single peer event as the individual most responsible for transmission. This was achieved by choosing the peer $\hat{u}$ with the strongest social influence on v at time t. That is,

$$\hat{u} = \arg\max_{u} \left[\varphi_{u,v}(t)\right] \tag{7.16}$$

Through this method uncovered the pattern of infections: each infection event is attributed to either the exogenous intensity or a single past infection event. We draw an edge from each infection event to all other infections that it spawns. We note that edges are directed forward in time, making cycles impossible, meaning that every connected component in this infection network is a tree. We referred to each tree in the forest of infections as a cascade.

The Timing of Infections and Co-Offending

Previous research suggests that co-offending represents strong and enduring relationships between individuals [502]. We therefore treated co-offending as evidence of an existing relationship between two individuals involved rather than as a point-in-time estimate of when that relationship formed, and accordingly generated static edges in the network representing that two individuals co-offended together at any time during the study period. Nonetheless, it is useful to evaluate the temporal relationship between when an individual is infected by an associate compared to when the two first co-offended together, to ensure that the typical timing of these two events supports this modeling decision. We considered all contagion events (as inferred in Section 7.2.6) between first-degree neighbors and found that 77.1% of all infected individuals had been co-arrested with their infector before being shot (Figure 7.8). Many of these infections occur in the immediate aftermath of being arrested with a recent victim. Another 10.8% of victims were shot in the year immediately preceding their first co-arrest with the infector. These results indicate that, even discounting prior research studying the close ties that generally exist between people before co-offending, our findings of contagion are not merely artifacts of the static network.

Causality in the Hawkes Model

The notion of causality has been the subject of many debates [388]. With this in mind, we should qualify the previous assignment of a single cause to certain gunshot victimizations.

In discussing the definition of causality, Ned Hall proposed the following thought experiment: "Suzy and Billy, two friends, both throw rocks at a bottle. Suzy is quicker, and consequently it is her rock, and not Billy's, that breaks the bottle. But Billy, though not as fast, is just as accurate: Had Suzy not thrown, or had her rock somehow been interrupted mid-flight, Billy's rock would have broken the bottle moments later" [209].

According to some interpretations of causality, within this scenario Suzy and Billy are jointly responsible for the bottle breaking: they were both throwing rocks at it, and the fact that Suzy's rock reached the bottle first is

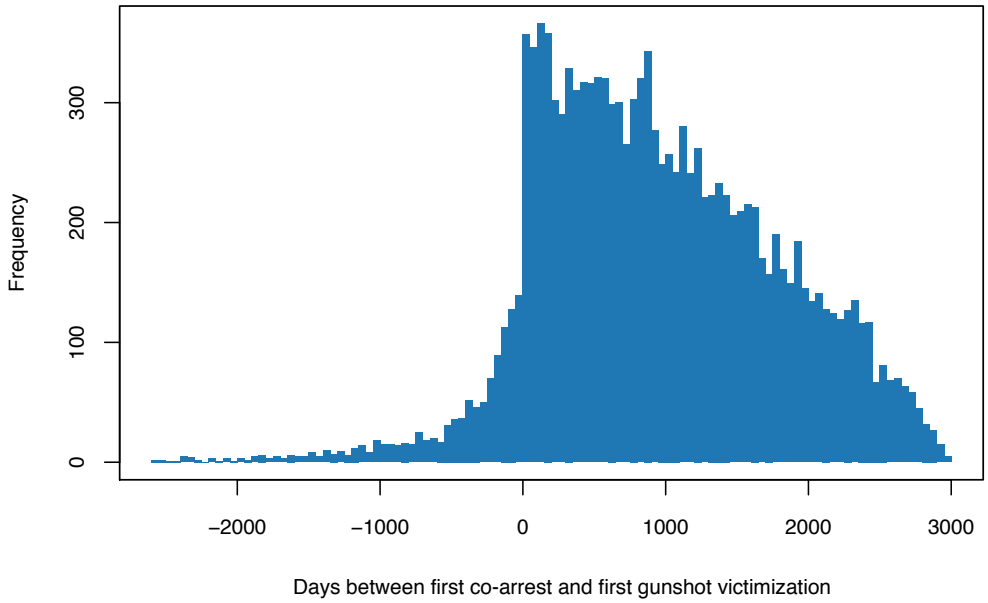

Figure 7.8: Temporal difference between co-offending and the transmission of gunshot victimization. The values here indicate the number of days between co-offending and being victimized, among cases where our model determined that a victim was infected by a first-degree neighbor. Positive values indicate that the victim was infected after having previously co-offending with the infector. As is clear from the figure, the majority of infections (77.1%) that we detected between first-degree neighbors occurred after the infector and victim had already been arrested together. Among the victims who were shot before co-offending with their infector, 47.5% co-offended with the infector within a year of being infected. These results, combined with previous research on the enduring nature of co-offending relationships [502], confirm the validity of modeling the contagion process over a static network.

coincidental. However, it also clear that Suzy's rock shattered the bottle. Even if we had not observed the rock that first hit the bottle, since Suzy was throwing rocks more quickly than Billy we could say that the rock that shattered the bottle was more likely to have been thrown by Suzy.

The Hawkes contagion model can be re-interpreted in light of this example: as they become infected, victims begin to "throw rocks" at their associates with a frequency that decreases over time. Being shot due to peer infection is equivalent, in this metaphor, to being hit by a rock thrown by an associate. Since we do not observe whose rock hits first, the only thing we can say for certain is that at the time of victimization an individual was subject to the combined throws of his or her previously-infected neighbors. This combined effect is expressed mathematically by the sum in Equation 7.1.

It is now clear which interpretation to give the cascades extracted in Section 7.2.6: it is a simplification where

we designate the cause for victimization to be the associate who was "throwing rocks" with the highest frequency at the time of infection. This simplification is acceptable in that this associate is the *most likely* to be the direct cause of infection. Nonetheless, based on another interpretation of causality we would instead consider the throws from every associate to be jointly the cause of victimization.

7.2.7 Model Evaluation: Predicting Victims

We compared the Hawkes contagion model with a traditional demographics model by evaluating how effectively each model predicts who will be shot on a given day. Given that social services must make targeted interventions with limited resources, predictions of gunshot victims are only actionable if they precisely identify a small population that faces the highest risk to be shot. With this in mind, the proper evaluation for any model is its ability to identify future victims as part of the population's highest-risk community [76, 345]. For this study, we define three "high-risk communities" as those people identified with the top 0.1%, 0.5%, and 1.0% of risk to become infected. These correspond to populations with 138, 691, and 1,382 individuals from the largest connected component, respectively.

We compared the predictive abilities of three different models:

Demographics Model: This model uses each person's demographic features and risk factors to predict who will be infected on a given day. We include all features available in our data, capturing many of the variables shown to be most critical in predicting gunshot victimization [480]. We label as infected all people who have been shot before that date and label all others as non-infected. We then perform a logistic regression over the entire population, using the formula $victim \sim sex + race + age + gang.member + gang.name + N.prior.arrests + neighborhood$ (while additional features would surely have been useful, we unfortunately did not have access to any variables beyond these). The resulting probabilities correspond to the background rate of the Hawkes contagion model and identify each person's risk to be shot.

Contagion Model: This model uses the social contagion element of the Hawkes model to identify who is

at most risk to become infected on a given day. It accounts for the network structure and infection history, but ignores all demographic and environmental attributes. Based on the observed pattern of gunshots, we measure each person's exposure to violence at a given time.

Combined Model: This model uses the results from the demographics and network models. We combine the risk scores from the other two models using a weighted sum, generating a fully specified Hawkes contagion model for the spread of gunshot violence through the co-offending network.

For every day of the study period, we executed all three models to predict each person's likelihood to be shot on that day. We then identified (based on the data) the people who were actually shot on the current day of the trial and noted their relative risk in the population of co-offenders according to each model. For each model, we ended up with the rankings of all the victims on the day they were shot. We compared the different models by measuring how often they select victims when identifying the network's high-risk population. An ideal model would identify each day's N victims as the individuals with the N highest levels of risk.

7.3 Results

7.3.1 Characteristics of the Network

Table 7.1 shows characteristics of the 138,163 people in the Network. Figure 7.9 provides a graphical representation of the Network, showing the relative locations of victims and non-victims. Individuals were on average 27 years old at the midpoint (in 2010) of the study, and predominantly male (82.0%) and Black (75.6%). According to police estimates, 26.2% were members of street gangs. Compared to non-victims, the victims of gunshots were 3.8 years younger (23.2 vs. 27.0 years) and more likely to be male (97.0% vs. 80.9%), Black (79.8% vs. 75.3%), and involved in a gang (52.3% vs. 24.3%). Consistent with prior research [385], gunshot victimization was highly concentrated within the network. Gunshot victims were socially close to other gunshot victims in the Network: 17.9% of victims' first-degree associates were also victims, compared to 9.8% for non-victims.

	LCC	Victims	Non-Victims
Demographics			
Number of people	138,163	9,773	128,390
Age at study midpoint	27.5	23.2	27.0
Percent male	82.0%	97.0%	80.9%
Percent Black	75.6%	79.8%	75.3%
Percent white/Hispanic	23.3%	19.5%	23.6%
Percent gang member	26.2%	52.3%	24.3%
Network characteristics			
Number of co-offenders (degree centrality)	6.1	10.2	5.7
Percent of neighbors who are victims (degree=1)	10.4%	17.9%	9.8%
Percent of neighbors who are victims (degree$\leq$2)	11.1%	15.9%	10.7%
Percent of neighbors who are victims (degree$\leq$3)	11.8%	14.9%	11.6%

Table 7.1: Characteristics of the 138,163 Individuals Arrested in Chicago from 2006 to 2014 and in the Largest Connected Component (LCC) of the Network. All comparisons between gunshot victims and non-victims were P<0.001 (P-values were calculated using the Welch Two Sample t-test).

This pattern was similar for second- and third-degree associates as well (see Table 7.1), indicating that there were clusters in the network with many victims and other parts with few victims.

7.3.2 Alternative Explanations

Our comparisons of simulations with the data show that homophily and confounding cannot fully explain the concentration of gunshot victims within the network (Figure 7.10). As reported in the main text, these pairs are shot on average 60 days closer together than the simulations can explain. We similarly found that the median time difference between victimizations is 75 days shorter in the data than in the simulations. We then evaluated how many pairs become victims within a specific, short period of time. While 7.6% of pairs in the data became victims within 30 days of one another (N=726), there were only 4.0% (3.7%-4.4% 95CI) such pairs in the simulations. Homophily and confounding, then, explained only 52.6% of the gunshot victimization that occurred between associates within 30 days. Similarly, 17% of pairs in the data became victims within 100 days of one another, compared to only 12% in the simulations. These results indicated that victims are clustered both temporally and

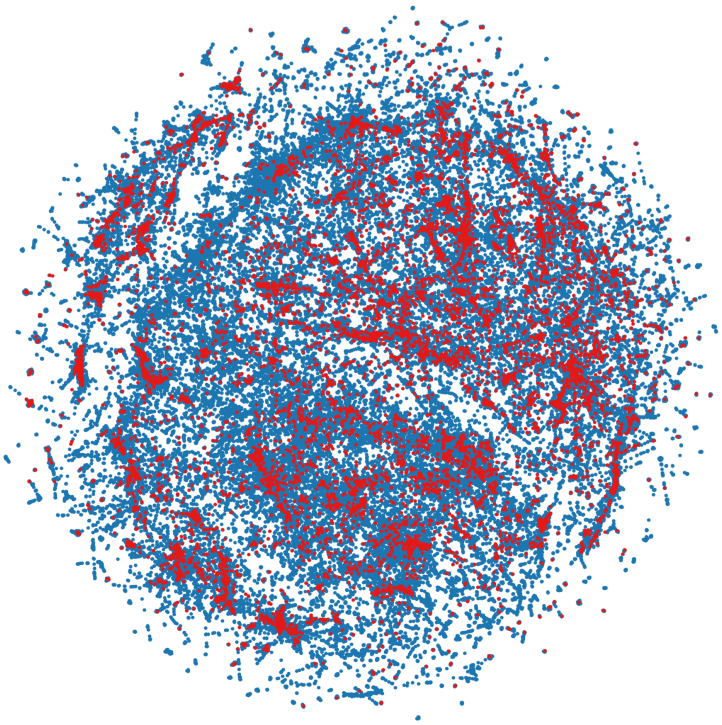

Figure 7.9: Graphic Representation of the Largest Connected Component of the Network. Each node represents a unique individual (N=138,163). Red nodes identify victims of a fatal or non-fatal gunshot injury (N=9,773); blue nodes represent people who were not victimized (N= 128,390). Data are from the Chicago Police Department as described in the text.

topologically in a manner that homophily and confounding cannot fully explain. This suggested that considering social contagion may help explain when and where victimizations occur in the network. We turn in the next section to modeling this social contagion directly.

7.3.3 Modeling Contagion

The distribution of the cascade sizes extracted from our dataset can be seen in Figure 7.11. Consistent with previous findings in related domains [80, 298], this distribution follows a power-law of exponent 1.8.

After calibrating our model to the data, we found that 63.1% (N=7,016) of the 11,123 gunshot victimizations in the Network during the study period were attributable to social contagion. This distribution was similar for both fatal (60.8%, N=829) and non-fatal injuries (62.6%, N=6,187). We found that 46% of infections came from first-degree neighbors, 41% came from second-degree neighbors, and the remaining 13% of infections

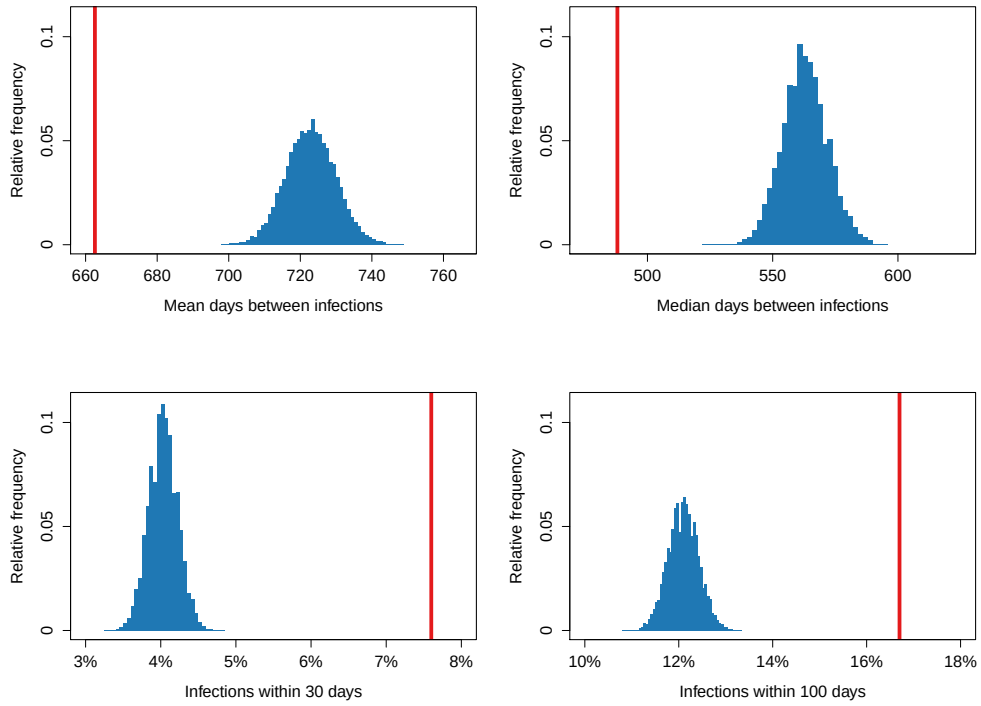

Figure 7.10: Results from 10,000 Monte Carlo simulations of the study period without any social contagion. These plots display the temporal relationships between infections for all pairs of first-degree neighbors where both people were gunshot victims during the study period. Vertical red lines represent the observed values from the data. Simulations based on homophily underestimate by a large margin how many pairs will be infected close together in time, and can explain only 52.6% of infections that occur within 30 days of each other. The mean time between infections is 60 days shorter in the data than in the simulations.

came from third-degree neighbors. Victims were shot on average 125 days after their infector (the person most responsible for the victim being exposed to gunshot violence), with a median time difference of 83 days.

From tracing gunshot victimization through the network, we detected 4,107 separate cascades (connected chains of infection through the network) ranging in size from cascades with a single person to a cascade involving 469 people, with 680 cascades involving multiple people and a mean cascade size of 2.7 people (Figure 7.11). Figure 7.12 depicts three representative cascades, containing 12 people, 34 people, and 64 people, all shot during the study period and showing the pathways of diffusion between individuals. These cascades visually reinforce how gunshot victimization spreads through a co-offending network, connecting individuals who initially had no connections to one another. They also help to explain the concentration of victims as shown in Table 7.1 and

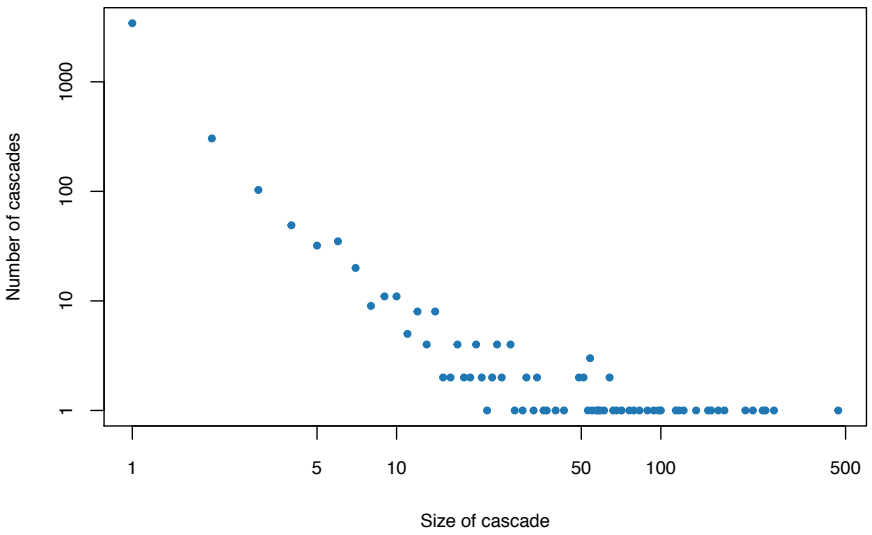

Figure 7.11: Distribution of cascade sizes found in the network. Cascade sizes ranged from 1 (N=3,427, 83.4%) to 469 (N=1), with a mean size of 2.7 people. The distribution follows a power law with scaling exponent 1.8.

Figure 7.9, since victimizations in one part of the Network generate further victimizations in that same region.

7.3.4 Predicting Victimization

Figure 7.13 shows a comparison of the three models to predict gunshot victimization: a model based on demographics, a model based on contagion, and a model based on both demographics and contagion. The contagion model outperformed the demographics model at estimating an individual's risk to be shot (Figure 7.13). Over the study period, the contagion model identified 5.3% of the Network's victims (N=589) among the 1.0% of the population it deemed highest-risk each day, compared to 4.3% (N=475) identified by the demographics model (24.0% increase). The combined model performed best, identifying 6.5% of victims (N=728) when selecting the 1.0% highest-risk population daily. Compared to the demographics model, across the three daily high-risk population sizes considered (0.1%, 0.5%, and 1.0%), the combined model correctly identified 71.7%, 65.5%, and 53.3% more victims, respectively.

Figure 7.14 plots the cumulative distribution function for each model. The contagion model outperformed the demographics model for the high-risk quarter of the population (identifying more than half of the victims

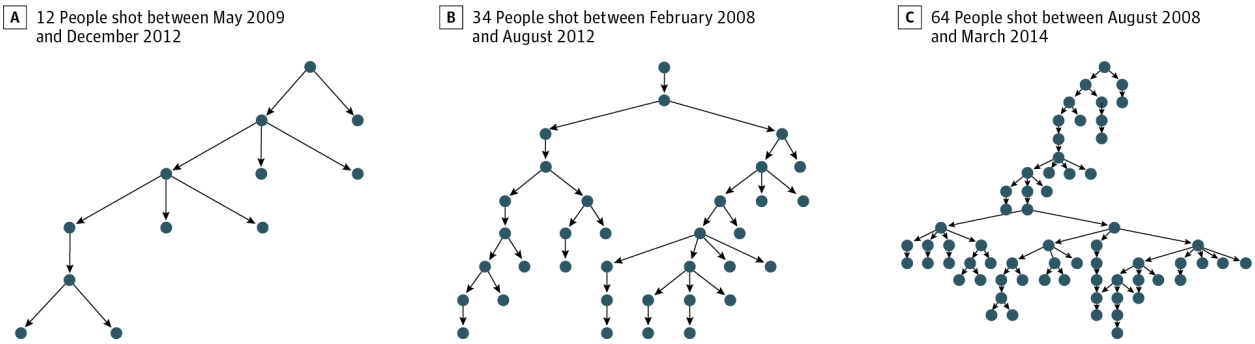

Figure 7.12: Three Cascades of Gunshot Victimization Inferred From The Study Period. Each edge (a line with small arrow showing direction) represents the transmission of gunshot victimization from one individual to another. The originators of each cascade are red; all other individuals infected as part of the cascade are blue. These cascades represent (A) 12 people shot between May 2009 and December 2012, (B) 34 people shot between February 2008 and August 2012, and (C) 64 people shot between August 2008 and March 2014.

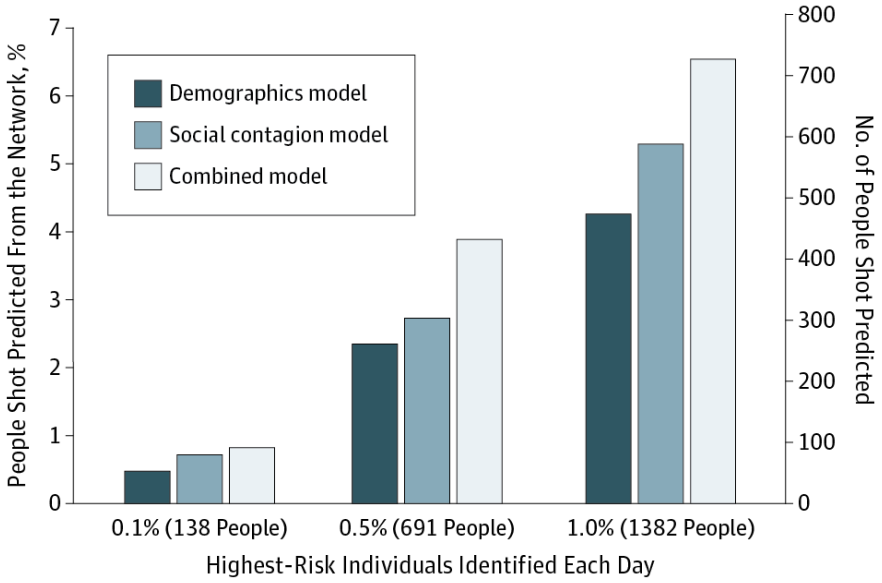

Figure 7.13: Predictions of Individual Risk of Gunshot Victimization Among High-Risk Populations. Comparison of the ability of the three models to identify victims as one of the highest-risk individuals in the Network on the day that the victim was shot; predictions for the 0.1%, 0.5%, and 1% of individuals at highest risk are shown.

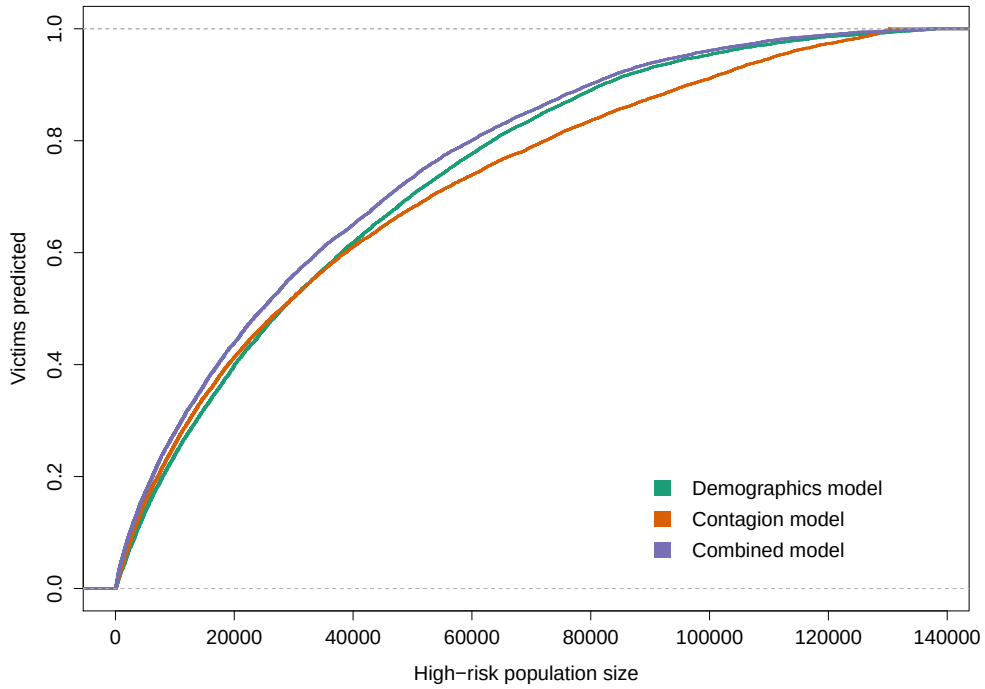

Figure 7.14: Cumulative distribution function for the demographics, contagion, and combined prediction models. The x-axis represents a population size and the y-axis reports what fraction of victims was within the high-risk population of that size. Among the highest-risk 20,000 people, for example, the demographics model identifies 39.9% of victims, the contagion model identifies 41.3%, and the combined model identifies 43.9%. Overall, the contagion model outperforms the demographics model for high-risk quarter of the population (identifying more than half of the victims in this group), while the demographics model outperforms the contagion model for the rest of the population. The combined model reaps the benefits of both models, and performs best across the entire distribution.

in this group), while the demographics model outperformed the contagion model for the rest of the population. The combined model reaped the benefits of both models, and performs best across the entire distribution. This shows that the contagion model is best equipped to predict future victims when focused on the portion of the population that faces the highest risk. Given that the goal of predicting victims is to provide social services with a small population for targeted interventions, the contagion and combined models are more effective than the traditional demographic model.

7.4 Discussion

Comparing levels of gun violence in the United States and its concentration within communities to an epidemic garners wide appeal but, scientifically, often stops at descriptive and spatial analyses. Whereas previous research has been cross-sectional, the current study advances understanding of gun violence by modeling it as social contagion and by directly tracking the contagion's spread. Our findings suggest not only that gunshot victimization concentrates within certain populations, but also that the diffusion of victimization follows an epidemic-like process of social contagion that is transmitted through networks by social interactions. Violence prevention efforts that account for contagion in addition to demographics to identify likely victims have the potential to prevent more shootings than efforts that focus only on demographics.

Our research suggests that a holistic public health approach to gun violence should be developed in at least two ways [227]. First, violence prevention efforts should consider the social dynamics of gun violence: tracing the spread of victimization through social networks could provide valuable information for public health and medical professionals, in addition to law enforcement, looking to intervene with the people and communities at highest risk. Given that public health and epidemiology are founded on studying pathways of transmission, approaches from these domains may readily extend to gun violence prevention efforts. For example, information on the timing and pathways of gunshot cascades might provide street outreach workers of campaigns such as Cure Violence (a violence prevention model, used in more than 50 U.S. cities, that draws on public health methods to mediate conflicts before they become violent) with a more accurate assessment of the people who would most benefit from their program [66]. Likewise, hospital-based violence intervention programs [81, 407] might follow such network models to extend their services beyond the emergency room to others within a social network who are also at risk of becoming gunshot victims.

Second, concerted efforts should focus on making gun violence prevention efforts victim- rather than offender-focused—namely, prioritizing the health and safety of those in harm's way. Although mounting evidence from

multiple cities suggests that small place-, group-, and network-based interventions can effectively reduce gun violence [52, 103, 270, 383], these network-based approaches have often relied heavily or solely on law enforcement activities. The individuals identified in our study are not just in contact with the criminal justice system—they are also deeply embedded within the public health, educational, housing, and other governmental systems. A fully realized victim-centered public health approach includes focused violence-reduction efforts that work in concert with efforts aimed at addressing the aggregate risk factors of gun violence—the conditions that create such networks in the first place or otherwise determine which individuals are in such networks (such as neighborhood disadvantage and failing schools).

Several limitations of our study should be noted. First, we lacked additional data that might have been relevant to understanding individual and neighborhood risk factors, such as substance abuse, employment, and police activity. Thus, our models may have underestimated the predictive ability of demographic and ecological risk factors. Second, although our descriptive findings of the Chicago co-offending network were quite similar to those from Boston and Newark [381, 382], additional research is needed to understand how city-specific factors like segregation, public housing policies, street gangs, and the availability of guns might influence both the structure of social networks and the transmission process related to gun violence within them. Finally, our study relied on a single behavioral tie, co-offending, and thus failed to capture other social ties (such as kinship, friendship, employment, and gang membership) that might also facilitate the contagion process or else protect individuals from infection. Specifically, we were unable to assess why some individuals in the social network—indeed, the vast majority—never became gunshot victims. Understanding resilience in networks is an important next step for research and practice, and future research should expand its focus on the types of networks that foster and abate the contagion of violence. Developing our understanding of resilience in networks might advance a preventative approach to mitigating the effects of gun violence that looks not simply to respond to shootings that have already happened, but to bolster networks that might inoculate from the potential for future shootings.

In conclusion, we analyzed administrative records to show how modeling gun violence as an epidemic that

spreads through social networks via interpersonal interactions can improve violence-prevention strategies and policies. Our results suggest that an epidemiological approach, modeled on public health interventions developed for other epidemics, can provide valuable information and insights to help abate gun violence within U.S. cities.

Part III

Reform

Chapter 8

The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness

8.1 Introduction

As described in the Introduction, the case for criminal justice risk assessments rests on a particular theory of change: first, that risk assessments will mitigate judicial biases by providing "objective" decisions about defendants, and second, that risk assessments will promote reform to (one aspect of) the criminal justice system by replacing discriminatory policies and reducing incarceration. Given the centrality of this theory of change to the use of risk assessments, evaluating risk assessments as an approach to criminal justice reform requires interrogating both underlying assumptions. This chapter connects the studies presented in the preceding chapters to the broader pursuit of criminal justice reform (particularly in the context of pretrial and sentencing).

This analysis requires, as a preliminary step, articulating principles with which to evaluate reform. This is particularly important given that the notion of "criminal justice reform" is itself contested. Criminal justice

reform refers broadly to the goal of eliminating or altering policies that lead to mass incarceration and racial injustice. However, there are divergent views about both the causes of and solutions for these challenges. For example, police reform efforts range from focusing on deficiencies in African American male culture (reforms require improving this culture) to the enduring presence of white supremacy and antiBlack racism (reforms require structural transformations to U.S. society) [62].

While it is expected that any reform effort will involve multiple visions, the rhetorical flexibility of "criminal justice reform" leads to a significant gap between the expansive change that "reform" suggests and the more minimal shifts that many reformers actually intend. As a result, criminal justice reform rhetoric is often both "superficial"—"most proposed 'reforms' would still leave the United States as the greatest incarcerator in the world"—and "deceptive"—many so-called reformers "obfuscate the difference between changes that will transform the system and tweaks that will curb only its most grotesque flourishes" [264].

This chapter evaluates reforms based on the extent to which they address the well-documented structural causes of carceral injustice. This is the emphasis articulated by the prison abolition movement, which draws on the slavery abolition movement [117, 332]. Formerly consigned to the outskirts of political discussion, abolition has been the subject of renewed attention among politicians, the legal academy, social movements, and the media [19, 416, 286]. Prison abolition promotes decarceration with the aim to ultimately create a world without prisons. Recognizing the violence inherent to confining people in cages and to controlling people's lives through force, abolitionists object to reforms that "render criminal law administration more humane, but fail to substitute alternative institutions or approaches to realize social order maintenance goals" [331]. Nor, however, do abolitionists intend to immediately close all prisons. Instead, abolition is a long-term project to develop "a constellation of alternative strategies and institutions, with the ultimate aim of removing the prison from the social and ideological landscapes of our society" [117]. This involves advocating to end practices such as capital punishment, the use of criminal records in determining access to housing and voting, and the militarization of police [416] and to create alternative practices such as transformative justice, democratic and holistic responses to violence, and

increasing resources for education and healthcare [333].

With the aim of structural decarceration in mind, this chapter interrogates the theory of change motivating risk assessments. First, building on sociotechnical approaches to objectivity, I demonstrate how the objectivity promised by risk assessments is a chimera: rather than *removing* discretion to create neutral and objective decisions, risk assessments *shift* discretion toward other people and decision points. Second, drawing on legal critiques of rights as tools for achieving just outcomes, I describe how risk assessments are an ill-advised tool for reducing the centrality and legitimacy of incarceration: risk assessments are indeterminate tools that provide no guarantee of reducing incarceration, are made ineffectual by their individualistic conceptions of risk and bias, and are likely to legitimize the structure and logic of criminal punishment. Rather than presenting a viable approach to decarceral criminal justice reform, risk assessments present a superficial solution that reinforces and perpetuates the exact carceral practices that require dismantling.

Risk assessments can, however, be reinterpreted to point toward more substantive criminal justice reform. A proper challenge to risk assessments requires not technical or procedural reforms, but an "epistemic reform" that provides a new interpretation of both risk assessments and the criminal justice system. Thus, having analyzed the impacts of risk assessments *within* the criminal justice system, I turn to questioning what risk assessments tell us *about* the criminal justice system. Returning to the "fairness" of risk assessments, I reinterpret recent results regarding the "impossibility of fairness" [82, 278] as an "incompatibility of equality." These impossibility results reflect not simply a tension between mathematical metrics of fairness, but instead indicate the fundamental conflict between approaches to achieving justice: *the impossibility of fairness mathematically proves that it is impossible to achieve substantive equality through mechanisms of formal equality*. This epistemic reform challenges the formalist, colorblind proceduralism at the heart of the criminal justice system and provides an escape from the seemingly impossible bind of fairness, exposing an expanded range of possibilities toward achieving criminal justice reform. Moreover, this analysis highlights the severe limitations of fairness as a method for evaluating the social impacts of algorithms, highlighting in particular how algorithmic fairness narrows the scope of judgments about justice

and how "fair" algorithms can reinforce discrimination.

8.2 Objectivity

Although objectivity is often used as a synonym for "science" and "truth," objectivity is only partially aligned with these terms [116, 395]. The meaning of objectivity comes most directly from its opposition to subjectivity: the goal behind objectivity "is to aspire to knowledge that bears no trace of the knower" [116]. Yet "[t]his ideal of mechanical objectivity, knowledge based completely on explicit rules, is never fully attainable" [395]. The practices followed to produce objectivity are themselves grounded in social norms about what kinds of knowledge are considered objective. These "methods for maximizing objectivism have no way of detecting values, interests, discursive resources, and ways of organizing the production of knowledge," meaning that "nothing in science can be protected from cultural influence" [214]. Thus, rather than producing knowledge that is truly free from the trace of any people, objectivity represents "knowledge produced in conformity with the prevailing standards of scientific practice as determined by the current judgements of the scientific community" [9].

Objectivity in the form of quantification plays a particularly important role in political contexts rife with distrust, in which officials facing external scrutiny need to depoliticize their actions by "making decisions without seeming to decide" [395]. Particularly in the United States, which is notable for its wariness of individual decision makers, "Techniques such as cost-benefit analysis and risk assessment make it easier to reassure critics within and outside government that policy decisions are being made in a rational, nonarbitrary manner" [249]. Nonetheless, "Study after study and commentary after commentary [have] called attention to the profoundly normative character of risk assessment, showing that it is a far from objective method: indeed, that it is a highly particular means of framing perceptions, narrowing analysis, erasing uncertainty, and defusing politics" [253].

Pretrial and sentencing risk assessments exemplify these attributes of objectivity. As concern about discrimination and mass incarceration intensifies, the criminal justice system faces heightened scrutiny. In order to

defuse these challenges and depoliticize their actions, criminal justice actors have turned to risk assessments. Practitioners such as probation officers have reported that risk assessments provide defensible grounds for their decisions, making them less vulnerable to criticism [210].

Rather than produce knowledge that lacks any trace of subjectivity, however, risk assessments produce information (and hence outcomes) that is embedded within political norms and institutional structures. Four aspects of risk assessments deserve particular attention as sites where their supposed objectivity breaks down and a great deal of hidden discretion is incorporated: defining risk, producing input data, setting thresholds, and responding to predictions. These sites of discretion exist in addition to the decisions that are inherent to the development of every machine learning model (such as selecting training data and model features [27, 194]) or are external to the risk assessment decision-making process itself (such as choosing what interventions should be pursued in response to risk). After this section describes these forms of discretion, the next section will analyze how such discretion hinders risk assessments as a tool for achieving substantive criminal justice reform.

8.2.1 Defining Risk

Risk assessments aim to predict risk, defined as the likelihood of crime. Pretrial risk assessments estimate the risk that a defendant will be rearrested before trial or will not appear for trial; sentencing and parole risk assessments estimate the risk that a defendant or inmate will recidivate. Such predictions typically consider a period of time ranging from six months to two years [326].

Forecasting crime while ignoring the impacts of incarceration causes risk assessments to overvalue incarceration.[1] Releasing someone by definition increases that person's likelihood to commit a crime in the near future. If crime risk is the primary criterion, then release will always appear to be adverse.

Yet there are many harms associated with incarceration. Pretrial detention significantly increases a defendant's likelihood to plead guilty, be convicted, and receive long prison sentences [133, 224, 314]. Time spent in prison

[1] In practice, risk assessments are based in data about arrests, which typically represents a racially biased measure of crime [146, 327].

is associated with negative outcomes including sexual abuse, disease, and severe declines in mental and physical well-being [263, 509]. After being released, former inmates face significant challenges in finding work (a barrier that is stronger for Blacks than whites) [379] and suffer disproportionately from depression, serious disease, and death [509]. The families and communities of incarcerated people also face severe hardships [176, 219, 506]. Moreover, because incarceration increases one's long-term propensity for crime, pretrial detention does not actually reduce future crime [133]. All told, a cost-benefit analysis found that "detention on the basis of 'risk' alone can lead to socially suboptimal outcomes" [518].

The emphasis on crime risk also causes risk assessments to absorb the highly racialized meaning of crime. As numerous scholars and lawyers have shown, the types of behaviors that society views with fear and chooses to punish are based in racial hierarchies, such that Blackness itself is criminalized [63, 213, 264, 354, 462] and "risk [is] a proxy for race" [213]. As such, risk assessments subsume the racialized concept of crime into a seemingly objective and empirical category that should guide decision-making.

8.2.2 Generating Input Data

Some risk assessments rely on information collected by a criminal justice practitioner (e.g., parole officer or social worker) via an interview with a defendant. For example, the widely-used COMPAS risk and needs assessment incorporates information from interviews that include questions such as "Is there much crime in your neighborhood?" [367]. Another risk assessment evaluates individuals along categories such as "Community Disorganization," "Anger Management Problems," and "Poor Compliance" [87].

Such questions and categories resist objective answers, turning these assessments into value-laden affairs in which white, Western, and middle-class standards are imposed on defendants [210, 325]. One's freedom can hinge on these assessments: in 2016, an inmate in New York was denied parole due to a rehabilitation coordinator answering "yes" to the question "Does this person appear to have notable disciplinary issues?" despite the inmate's lack of a single disciplinary infraction over the prior decade [507].

Recognizing that their evaluations influence the calculations and recommendations of risk assessments, many criminal justice practitioners exercise "considerable discretion" in collecting and interpreting information to produce what they see as the appropriate final score [210]. One study found that practitioners ignored or downplayed criminogenic factors in order to produce low risk designations when evaluating minorities who had committed low-level offenses, but interpreted information so as to produce high risk scores when evaluating sexual or violent offenders [210].

8.2.3 Setting Thresholds

Once someone's risk has been predicted, risk assessments turn the forecasted probability into categories (e.g., low/medium/high [17]) and number ranges (e.g., 1-5 [311]) to be presented to judges. Notably, no prominent risk assessment directly presents probabilities [86] or follows the "intuitive interpretation" [146] of dividing categories across the spectrum of risk (e.g., "low risk" corresponds to 0-33% risk). A related approach is to define risk categories across population percentile (e.g., COMPAS divides the population into ten equal-sized groups, assigning each a score from 1-10 [366]).

In most cases, therefore, the thresholds that determine labels such as "high risk" and recommendations such as "detain" are based in normative judgments about the tradeoffs between reducing incarceration and reducing crime. Jurisdictions implementing the PSA, for example, determine how to define the ranges of low, moderate, and high risk [463]. Although there may be benefits to adapting risk assessments to the local context, doing so introduces a new form of discretion: there is no objective guide for what certain level of risk warrants release or detention. Across risk assessments, the probabilities corresponding to the highest risk categories vary widely and can refer to rearrest rates as low as 3.8% [22, 326]. In turn, public officials often do not actually know how the categories that risk assessments present translate to probabilities of recidivism or failure to appear [264, 281].

These scores and thresholds can have significant impacts on the outcomes of cases. Many jurisdictions directly tie recommendations to the categories defined in the risk assessment [292, 449]. In Kentucky, for instance, the

mandatory use of a pretrial risk assessment led to increases in release for low and medium risk defendants and a decrease in release for high risk defendants [464].

Even if a recommendation threshold is set at the outset of reform to promote high levels of pretrial release, it can later be altered to reduce pretrial release. In New Jersey, several defendants accused of certain gun charges were released before trial and then rearrested; the Attorney General's office then pressured the courts to alter the risk assessment so that it would recommend detention for every defendant arrested for those same gun charges, regardless of that person's predicted risk [235, 440]. New Jersey soon expanded its detain recommendations to a larger number of offenses [394]. Similarly, in 2017, the United States Immigration and Customs Enforcement (ICE) altered its pretrial Risk Classification Assessment so that it would recommend "detain" in every case [422].

8.2.4 Responding to Predictions

Regardless of how they present predictions, risk assessments typically play a role of decision-making aid rather than final arbiter: they provide information and recommendations to judges but do not dictate the decisions made. Thus, although a common goal behind risk assessments is to eliminate the subjective biases of judges [495, 99, 406, 340, 245, 460, 471, 474], risk assessment implementations allow judges to decide how to respond to the information and recommendations provided.

Many judges use this discretion to ignore risk assessments or to use them in selective ways. In both Kentucky and Virginia, risk assessments failed to produce significant and lasting reductions in pretrial detention because judges tended to override recommendations suggesting release [464, 465]. Judges in Cook County, Illinois diverged from the pretrial risk assessment 85% of the time, releasing defendants at drastically lower numbers than recommended [319]. A juvenile risk assessment faced similar issues: judges frequently overrode the risk assessment when it recommended release, but rarely when it recommended incarceration, leading to a dramatic and "chronic" increase in detention [461]. Similar patterns have been observed in Santa Cruz and Alameda County, California [503].

When they do use risk assessments, judicial decisions are rife with biases. Two experimental studies found that people are more strongly swayed by a risk assessment's suggestion to increase estimates of crime risk when evaluating Black defendants compared to white defendants (see Chapters 3 and 4). Judges in Broward County, Florida have penalized Black defendants more harshly than white defendants for being just above the thresholds for medium and high risk [107]. Judicial decisions made with a risk assessment in Kentucky similarly increased racial disparities in pretrial outcomes [8].

It is clear that the first assumption behind risk assessments—that they replace biased discretion with neutral objectivity—does not hold up to scrutiny. Despite being hailed as "objective," risk assessments shift discretion to different people and places rather than eliminate discretion altogether. Yet the presence of subjective judgment is not itself dispositive as an argument against risk assessments. For if the objectivity sought in risk assessment discourse is impossible, then any reform will rest, to some degree, on discretion. It is therefore necessary to turn to the second assumption motivating risk assessments and evaluate, with these subjectivities in mind, whether risk assessments can spur criminal justice reform.

8.3 Criminal Justice Reform

Although advocates tend to assume that risk assessments will promote reform in pretrial and sentencing adjudication [495, 217, 393], altering decision-making procedures to promote fairness and objectivity does not necessarily reduce incarceration and racial discrimination. Sentencing reform offers a striking case of how the "well-intentioned pursuit of administrative perfection" characteristic of twentieth century civil rights reforms "ultimately accelerated carceral state development" [354]. In 1984, concerned about the racial disparities produced by the judicial discretion to set criminal sentences, Congress passed the Sentencing Reform Act, creating mandatory sentencing guidelines tied to the characteristics of the offender and the offense [315]. This system was designed to constrain judicial discretion and thereby "provide certainty and fairness in meeting the purposes

of sentencing" [489]. The reform failed to have the intended impacts, however. The guidelines "set in motion dramatic changes in day-to-day federal criminal justice operations, largely by shifting a massive amount of discretionary power from judges to prosecutors" [315]. The result was a "punitive explosion" that increased both incarceration and racial disparities [315].

Similar reforms throughout U.S. history have centered on the expansion of rights as a mechanism to promote fair procedures. These rights-based reforms often did not actually notably improve outcomes: for instance, schools remained segregated and unequal well after the Supreme Court deemed school segregation unconstitutional in *Brown v. Board of Education* [483]). U.S. legal scholars in the twentieth century therefore developed the "critique of rights"—a critique of rights-based reforms and discourse in mainstream legal thought. Advanced by scholars such as Duncan Kennedy [267] and Mark Tushnet [482, 483], the critique of rights revolves around five assertions: 1) Rights are less effective at spurring progressive social change than commonly assumed, 2) The impacts of rights are indeterminate, 3) The discourse of rights abstracts away the power imbalances that create injustice, 4) The individualistic discourse of rights prioritizes individual freedom over social solidarity and community well-being, and 5) Rights can impede democracy by reinforcing undemocratic relationships and institutions [78].

Today's appeals to risk assessments mirror historical appeals to rights: like rights reforms such as the right to a lawyer, the introduction of risk assessments into bail and sentencing is intended to produce a fair and neutral process for criminal defendants [495, 217, 361]. This suggests that risk assessments should be interrogated against the critique of rights. Doing so, I show that risk assessments suffer from the same core limitations as rights: they are indeterminate, individualistic, and legitimizing.

8.3.1 Indeterminate

Although risk assessments are often hailed as objective, a great deal of subjective judgment resides under the surface of these tools. This discretion can dramatically alter the use and impacts of risk assessments. In this

sense, risk assessments are indeterminate: the adoption of risk assessments provides little guarantee that the intended social impacts will be realized.

Indeterminacy is a common feature of decision-making processes grounded in rules and procedures [483]. Procedural reforms often fail to generate the intended outcomes because they use technical means to achieve normative ends. Achieving the desired outcomes requires a particular use of the tool or process, yet nothing about the procedures guarantee that such use will arise in practice. As noted in the critique of rights, the adoption of a progressive law provides little guarantee of the political outcomes seemingly connected to that law; instead, broader social circumstances largely dictate how that law will be wielded, interpreted, and applied. And "if circumstances change, the 'rule' could be eroded or [even] interpreted to support anti-progressive change" [483].

Risk assessments are unreliable as tools for reducing incarceration because they depend on the social and political circumstances of their use. Risk assessments are embedded in the criminal justice system, in which the structural and political incentives largely favor punitive and carceral policies [10, 61, 63, 465]. Thus, to the extent that the types of subjectivity described in Section 8.2 manifest in risk assessments, such discretion typically resists decarceral goals. Definitions of risk emphasize incarceration as a way to reduce crime while ignoring the significant harms of incarceration. Interviews and evaluations allow white and middle-class assumptions (which typically associate Blackness with crime, aggression, and a lack of innocence [178, 182, 358, 408]) to influence judgments about defendants and inmates. The practice of defining thresholds allows for people with low probabilities of rearrest to be labeled "high risk" and for recommendations to be altered to reduce how many people are released. Judicial responses to risk assessments exacerbate racial disparities and diminish release rates.

These forms of discretion make the impacts of risk assessments brittle and prone to political capture. Achieving decarceral outcomes through risk assessments requires particular behaviors and circumstances which the criminal justice system is generally not amenable to. As a result of this indeterminacy, risk assessments provide no guarantee of reducing incarceration and in fact are often wielded in ways that resist decarceral outcomes. Yet because of the discourse that positions risk assessments as a tool for reform, even ineffective implementations

may enhance perceptions of fairness and reduce the political will for more systemic changes.

8.3.2 Individualistic

Risk assessments are based on individualistic conceptions of both risk and bias that lead to individualistic and ineffectual remedies for racial discrimination and mass incarceration.

Risk assessments treat risk at the level of individuals, defining risk in terms of someone's likelihood to be arrested in the future. This approach treats risk as a measure of difference across individuals—an objective and static fact of identity—rather than as a social category defined through social norms (what is considered a crime) and relations (why certain people commit and are punished for those crimes).

Although numerous social markers of difference are accepted as "intrinsic" and "natural," many of these categories emerge from social arrangements that imbue those comparisons with meaning and importance [341]. In particular, "difference" becomes salient when "a more powerful group assigns meaning to a trait in order to express and consolidate power" [341]. For example, "[w]omen are compared with the unstated norm of men, 'minority' races with whites, [and] handicapped persons with the able-bodied" [341]. Addressing difference equitably requires not providing special treatment (whether ameliorative or punitive) to "different" individuals, but altering the relationships and institutions that structure these categories [341].

Risk assessments focus on individual-level risk, leading them to suggest individual-level interventions. Calculating each person's risk differentiates risk factors across members within a population, but obscures the structural factors that shape the distribution of risk itself [405]. In other words, risk assessments make legible the idea of high-risk *individuals* rather than high-risk *populations*. As a result, risk assessments justify individualistic responses: most notably, incarcerating high-risk people. Yet it is precisely population-level reforms such as improving access to housing, healthcare, and employment that are most likely to reduce crime risk and improve well-being across the population [211, 225, 391, 431, 432].

Because risk assessments focus on individuals, they can entrench historical injustice by failing to recognize

changing social circumstances. Risk assessments (as with all machine learning) assume that population characteristics are constant, such that factors producing certain outcomes in the past will produce those outcomes at the same rates in the future. Even if jurisdictions enacted reforms that reduce crime, risk assessments would be blind to these new circumstances. In turn, risk assessments would overestimate crime and recommend incarceration for individuals whose crime risk has decreased. Following interventions such as text messages that remind defendants to appear in court, risk assessments have produced "zombie predictions" that overestimate risk because they fail to account for the risk-reducing benefits of these reforms [281]. And because incarceration increases the likelihood of crime after someone is released [111, 120, 497], these false positive predictions will exacerbate the cycle of recidivism and incarceration that risk assessments are meant to remedy.

Risk assessments suffer from a similarly individualistic approach to bias: they diagnose bias as a behavior exhibited by individuals, typically due to implicit bias. Risk assessments are therefore designed to replace the discretion of biased judges with "objective" algorithmic predictions [495, 99, 406, 340, 245, 460, 471, 474].

Yet this emphasis on the bias of individuals overlooks the policies and institutions that structure racial hierarchies. Discrimination and oppression are produced not simply by people making biased judgments, but through laws and institutions that systematically benefit one group over another [10, 264]. Diagnosing discrimination as the product of discretion and bias "displace[s] questions of justice onto the more manageable, measurable issues of system function" [354], thus "obscur[ing] the larger structural aspects of racism" and "draining attention and resources away from other approaches to framing and addressing racism" [258].

By focusing on judicial decisions as the source of discrimination, risk assessments shroud the social structures and power dynamics behind racial discrimination. They obscure the need to transform policies and institutions in order to achieve racial equity, instead suggesting that discrimination can be remedied by altering decision-making procedures. Attempts to address racial oppression that focus solely on the bias of individual decision makers serve to legitimize and reinforce that oppression.

8.3.3 Legitimizing

Despite being implemented under the banner of criminal justice reform, risk assessments naturalize and legitimize carceral logics (e.g., risky defendants should be held before trial) and practices (e.g., determining which defendants are "risky").

Across domains, reforms that address the salient aspects of an injustice rather than the underlying causes and conditions of that injustice can legitimize those underlying structures. For instance, efforts to eradicate war crimes such as torture without challenging war itself have "tolerated the normalization of perpetual, if more sanitary, war" [352]. Closer to criminal justice reform, diversity and implicit bias trainings present a notable example of how reforms aimed at preventing discrimination can legitimize social arrangements that produce inequality. Numerous studies have found these trainings to be ineffective at improving diversity or reducing bias [258]. Instead, by creating "an illusion of fairness" that "legitimize[s] existing social arrangements" [260], the "formal bureaucratic procedures may reproduce inequality rather than eradicating it" [271].

Individualistic and procedural reforms are particularly prone to legitimization. When it comes to legal rights, "progressive victories are likely to be short-term only; in the longer run the individualism of rights-rhetoric will stabilize existing social relations rather than transform them" [483]. This observation, that winning a legal battle can rely on principles (such as individualism) that hinder long-term efforts for structural transformation, is known as "losing by winning" [483]. With regard to criminal rights (such as the guarantee that every criminal defendant be provided with an attorney), "procedural rights may be especially prone to legitimate the status quo, because 'fair' process masks unjust substantive outcomes and makes those outcomes seem more legitimate" [65]. The enactment of such rights "makes it more work—and thus more difficult—to make economic and racial critiques of criminal justice" [65].

Risk assessments exemplify an individualistic and procedural reform as well as the limits of this approach. Risk assessments focus on decision-making procedures: their primary concern is not that incarcerating people

is wrong, but that decisions about which individuals to incarcerate should be reached more empirically and objectively. This represents a narrow vision of reform, one that attempts to measure risk without bias or error while upholding the notion that incarceration is an appropriate response to "high-risk" individuals. In other words, risk assessments focus reform efforts on decisions about individuals while overlooking the structures shaping that decision, who is subject to it, and what its impacts are. Although presented under the banner of reform, this type of "[a]dministrative tinkering does not confront the damning features of the American carceral state, its scale and its racial concentration" [354]. Instead, by tweaking surface-level decisions and providing them with a semblance of neutrality and fairness, risk assessments are likely to sanitize, legitimize, and perpetuate the criminal justice system's carceral and racist structure. From the perspective of decarceration and racial justice, the enactment of risk assessments represents a clear example of "losing by winning."

This process of legitimation can be seen most clearly with regard to preventative detention (detaining a criminal defendant before trial due to concerns about crime risk). The practice was not deemed constitutional until the 1987 U.S. Supreme Court case *United States v. Salerno* [26, 281, 518]. Yet today the practice of preventative detention—which Supreme Court Justices Marshall and Brennan deemed "incompatible with the fundamental human rights protected by our Constitution" [492]—is being legitimized as a central aspect of "modern" [361] and "smart" [494] criminal justice reforms based on risk assessments. Through such logic, the use of risk assessments as tools for reform "conced[es] Salerno" and "ratifies recent erosions of the fundamental rights of the accused" [281].

These three attributes of risk assessments—indeterminate, individualistic, and legitimizing—demonstrate the flaws of the assumption that risk assessments will promote criminal justice reform (at least with regard to any notion of reform that involves reducing the centrality of punishment and incarceration). These tools are poorly suited to the task of combatting carceral practices and logics. Despite being presented as a valuable mechanism for racial justice, risk assessments are akin to the many components of criminal justice reform today that are oriented around "the margins of the problem without confronting the structural issues at its heart" [264].

Thus far I have shown that the theory of change behind risk assessments is deficient: neither of the two core assumptions, regarding objectivity and reform, withstand close inspection. The question that remains is what this suggests for pretrial and sentencing reform efforts: How can risk assessments be challenged in a manner that facilitates a path toward more systemic reform?

8.4 Epistemic Reform

The movement for risk assessments derives not simply from the presence of particular technologies (i.e., big data and machine learning), but from a particular understanding of social challenges as technological in nature and amenable to technological solutions. As pressure mounts for criminal justice reform in an era of "technological solutionism" [349], "technochauvinism" [56], and "tech goggles" [194], what has emerged is a "sociotechnical imaginary"—a collective vision of a desirable future attainable through technology [254]—that casts criminal justice adjudication as prediction tasks, ones that algorithms can perform better than humans. Holding together these imaginaries and technologies is "co-production," which describes how "the ways in which we know and represent the world (both nature and society) are inseparable from the ways in which we choose to live in it" [250]. Through co-production, it is often new technological *discourses* rather than new technological *artifacts* that provide a sense of order in the face of instability [251]. Yet these discourses, however secure and widespread they may appear, are not static: altering forms of knowledge "can function as strategic resources in the ongoing negotiation of social order" [318].

This emphasis on discourses in addition to artifacts can inform the appropriate responses to the false promises of risk assessments. The dangers of risk assessments are not the result of poor implementation, but are instead inherent to the sociotechnical imaginary that treats criminal justice adjudication as a set of prediction problems. Under this framing, attempts to generate "better" (i.e., fairer and more accurate) risk assessments are unlikely to reduce these tools' fundamental harms. Rather than calling for unbiased risk assessments, then, a more fruitful

path to diminishing carceral logics and practices is to present an "epistemic challenge" [500] to the sociotechnical imaginary around risk assessments. Such an "epistemic reform" can shift our focus from evaluating risk assessments through the lens of the criminal justice system to evaluating the criminal justice system through the lens of risk assessments. Doing so can point the way toward more effective criminal justice reforms.

8.4.1 Reinterpreting the "Impossibility of Fairness"

Although it is common to discuss risk assessments and judges using the same language of bias—and even to directly compare their biases [35, 474]—"bias" has distinct meanings across these two contexts. The bias of a judge speaks to something individual: the implicit and explicit biases that influence a specific person's decisions. The "bias" of a risk assessment, on the other hand, speaks to something structural: the ways in which different groups of people are systematically filtered to different outcomes.

To understand this distinction, it is necessary to distinguish between two causes of algorithmic "bias":

1. Human Bias: The first form of "bias" occurs when an algorithm is trained on the decisions of biased humans—a type of "garbage in, garbage out." For instance, a risk assessment would be subject to Human Bias if its training data overestimates the recidivism rates of Black defendants due to over-policing in Black neighborhoods. Because this algorithm would be learning to reproduce human biases, it seems appropriate to refer to its decisions as "biased" and to make the comparison with human bias.

2. Population Inequity: The second form of "bias" occurs when an algorithm is trained on population-level disparities—a type of "inequity in, inequity out." For instance, a risk assessment would be subject to Population Inequity if its training data reflects (beyond any distortion from Human Bias) that Black defendants are more likely than white defendants to recidivate. Because this algorithm would be learning to reproduce social outcomes that are the product of historical oppression, its discrimination is not akin to the bias of human decision makers.

Failing to distinguish Human Bias from Population Inequity can hinder efforts to understand and reduce algorithmic discrimination.[2] Population Inequity is most directly related not to the biases of judges or other people, but to "the racial inequality inherent in *all* crime prediction in a racially unequal world" [327].

To see this challenge of making fair predictions in an unequal society, consider the recent statistical results regarding the "impossibility of fairness" [82, 278]. The results concern two metrics for evaluating fairness. The first metric is calibration, which states that predictions of risk should reflect the same underlying level of risk across groups (i.e., 50% risk should mean a 50% chance of rearrest whether the defendant is Black or white).[3] Calibration is akin to colorblindness. The second metric is error rate balance, which states that false positive and false negative rates should be equal across groups. Given these two metrics, the impossibility of fairness shows that if two groups have different rates of an outcomes, then it is impossible for predictions about those groups to both be calibrated and have balanced errors. In the context of risk assessments, this means that given higher crime rates among Black defendants than white defendants, it is impossible for a risk assessment to make calibrated predictions of risk without having a higher false positive rate (and lower false negative rate) for Black defendants.

From this perspective, risk assessments appear to be situated within an "impossible" set of tradeoffs [200]. In turn, the impossibility result is often interpreted as a defense of calibrated decision-making. When ProPublica demonstrated COMPAS' disproportionate false positive rates for Black defendants [17], Northpointe (the company, now known as Equivant, that created COMPAS) refuted that higher recidivism rates among Blacks explained the disparity and thereby absolved them from accusations of racial bias. They wrote, "This pattern does not show evidence of bias, but rather is a natural consequence of using unbiased scoring rules for groups that happen to have different distributions of scores" [130]. Other scholars have similarly pointed to the incom-

[2]Another paper has made a similar distinction, between 1) "racial distortions in past-crime data relative to crime rates" and 2) "a difference in crime rates by race" [327]. The two phenomena can also coexist: they are distinct but not mutually exclusive.

[3]A related measure is predictive parity, which states that the outcome rates among people labeled "high risk" should be the same across groups.

patibility of fairness metrics to dispel claims that algorithms are discriminatory [35, 99, 474].

Yet the problem of discrimination is not so neatly resolved by reference to the underlying base rates: the disparities in these population-level statistics are *themselves* the product of discrimination. African Americans do not just "happen to have different distributions of scores"—Blackness itself is criminalized [63, 213, 264, 354, 462] and Blacks have been subjected to myriad forms of oppression (including redlining and segregation [424], the war on drugs [10, 264], and severe underfunding of schools [147]) that contribute to increasing crime [284, 308, 420, 430].

Notions of fairness in risk assessments generally fail to consider such context, however. In both research and practice, calibration is the typical instantiation of fairness [101, 130, 327]. Yet calibration strives for accurate predictions of risk, regardless of the factors structuring that risk. Risk assessments thus overlook the social conditions behind racial disparities, striving to accurately identify risk without interrogating whether that notion of risk is appropriate, why some people have high levels of risk, or whether incarceration is an appropriate response for high-risk people. Rather than being blind to color, calibrated risk assessments are blind to structural oppression.

Consider the gold standard: a hypothetical risk assessment that predicts with perfect accuracy whether each person will recidivate.[4] Such a risk assessment would satisfy all three metrics of fairness that are typically in tension [278]. The impossibility would disappear. Yet this risk assessment would *still* disproportionately label Blacks as "high risk" compared to whites—not because of Human Bias, but because of Population Inequity: due to discrimination and the racialized meaning of "crime" and "risk" [63, 213, 264, 354], African Americans *are* empirically at higher risk to commit crimes [96, 431, 452, 487]. In other words, because "[r]acism is not a mistake, not a matter of episodic, irrational behavior" [121], eliminating inaccurate predictions will not eliminate racist predictions.

Herein lies the danger of overlooking Population Inequity: accounting only for Human Bias, even with a

[4]For the sake of this example, suppose that the training data and outcomes reflect an accurate and unbiased measure of crime (i.e., there is no Human Bias).

"perfect" risk assessment, would still subject Blacks to higher rates of incarceration than whites. This "fair" algorithm launders the products of historical discrimination into neutral and empirical facts, in turn reinforcing this discrimination by punishing African Americans for having been subjected to such criminogenic circumstances in the first place.

This conflict in algorithmic fairness between Human Bias and Population Inequity speaks to a more fundamental tension between notions of equality: formal equality and substantive equality. This tension runs throughout debates in areas ranging from equality of opportunity [165] to antidiscrimination [109] to big data [27]. Formal equality emphasizes equal treatment or equal process: similar people should be treated similarly. Substantive equality emphasizes equal outcomes: groups should obtain similar outcomes, even if that requires accounting for different social conditions between groups. In the U.S. legal context, disparate treatment is grounded in notions of formal equality (or anticlassification) while disparate impact is grounded in notions of substantive equality (or antisubordination).

By ensuring that individuals who have similar levels of risk are treated similarly, calibration expresses the logic of formal equality.[5] In this sense, calibration aims to account for Human Bias: it strives for predictions that reflect one's actual level of risk, untainted by distortions. Alternatively, by ensuring that groups are similarly affected by false predictions, error rate balance expresses the logic of substantive equality [327]. In this sense, error rate balance aims to account for Population Inequity: it strives for risk predictions that do not disproportionately harm one group more than another, regardless of the underlying distributions of risk.

With these parallels in mind, the epistemic reform becomes possible: the "impossibility of fairness" can be reinterpreted as an "incompatibility of equality." Because calibration is a measure of formal equality and error rate balance is a measure of substantive equality, the impossibility result can be restated as a tradeoff between formal and substantive equality: *the impossibility of fairness mathematically proves that, in an unequal society, decisions based*

[5]Although Mayson characterizes calibration as a disparate impact metric, I argue that it more closely aligns with the disparate treatment logic of ensuring that people with the same risk receive the same score (an equivalence that Mayson acknowledges) [327].

in formal equality are guaranteed to produce substantive inequality. Although the impossibility of fairness is typically taken to indicate that disparate outcomes are the mere byproduct of fair risk assessments [35, 99, 130, 474], this reframing highlights the opposite: disparate outcomes are the inevitable product of colorblind risk assessments in an unequal society.

Notably, it is precisely the desire for objectivity that grounds risk assessments in formal equality and makes them unable to generate substantive equality. Dominant notions of racial equality based in colorblindness developed from a desire for neutrality and objectivity, in direct opposition to more radical calls for racial justice from the Black nationalist movement [389]. Because colorblindness entails "the refusal to acknowledge the causes and consequences of enduring racial stratification" [354], it "creates and maintains racial hierarchy much as earlier systems of control did" [10]. Thus, just as the law "will most reinforce existing distributions of power" when it is "most ruthlessly neutral" [316], risk assessments will most entrench racial injustice when they are most (seemingly) objective.

8.4.2 Implications for Criminal Justice Reform

Statistical arguments that articulate these tensions between formal and substantive equality can challenge fundamental inequities in the criminal justice system. The Supreme Court confronted this issue in the 1987 case *McCleskey v. Kemp*, in which Warren McCleskey, an African American convicted of killing a white police officer, was sentenced to the death penalty in Georgia [491]. McCleskey challenged this verdict with statistical evidence of structural inequality: the death penalty was disproportionately applied in murder cases with Black defendants and white victims [25].

Despite this evidence, the Supreme Court affirmed the death penalty ruling. It argued that the statistical evidence failed to demonstrate deliberate racial bias in McCleskey's case. In the majority opinion, Justice Lewis Powell wrote, "a defendant who alleges an equal protection violation has the burden of proving the existence of purposeful discrimination. […] McCleskey must prove that the decisionmakers in *his* case acted with discrimina-

tory purpose" [491]. Powell provides a formal equality analysis: the outcome is legitimate as long as McCleskey was not subject to intentional discrimination.

Powell further justified this outcome by arguing that acknowledging substantive inequality in the face of formal equality would cause the entire structure of criminal law to crumble. Recognizing that "McCleskey challenges decisions at the heart of the […] criminal justice system," he wrote,

> In its broadest form, McCleskey's claim of discrimination extends to every actor in the Georgia capital sentencing process, from the prosecutor who sought the death penalty and the jury that imposed the sentence, to the State itself that enacted the capital punishment statute and allows it to remain in effect despite its allegedly discriminatory application. We agree with the Court of Appeals, and every other court that has considered such a challenge, that this claim must fail. [491]

The epistemic reform regarding risk assessments can embolden the discourse that Justice Powell recognized as an existential threat to the criminal justice system. Reinforcing the work of other scholars who have articulated the tensions between formal and substantive equality with regard to race and sex [109, 122, 317, 354, 389], the impossibility of fairness provides a mathematical proof of the inherent conflict between formal equality procedures and substantive equality outcomes in an unequal society. Failing to acknowledge the legacy of historical oppression will allow even "fair" risk assessments to perpetuate racial inequity. As Justice William Brennan remarked in his dissent in *McCleskey*, "we remain imprisoned by the past as long as we deny its influence in the present" [491].

Furthermore, the systematic nature of risk assessments may allow the incompatibility of equality to carry more force than the statistical evidence in *McCleskey*. In *McCleskey*, the Supreme Court argued that some variation in the outcomes of similar cases results from the "discretion [that] is essential to the criminal justice process" [491]. Risk assessments are specifically designed to replace judicial discretion with standardized objectivity, however. Moreover, algorithmic discrimination reflects not the bias of an individual but the systematic filtering of different groups into disparate outcomes. To the extent that judgments are standardized by risk assessments, then,

statistical evidence of racial disparities could become increasingly difficult to defend on procedural grounds of discretion and could instead be recognized as reflecting structural discrimination.

Such evidence on its own will not function as a "deus ex data" that prompts a restructuring of the criminal justice system. One lesson to be learned from *McCleskey* is that social scientific evidence may do very little to persuade courts to accept claims of discrimination [64]. Indeed, despite ProPublica's evidence that COMPAS disproportionately labeled Black defendants with false positive predictions of recidivism, the Wisconsin Supreme Court upheld the use of COMPAS at sentencing in *State v. Loomis* [514].

Achieving decarceral reform therefore requires emphasizing the *interpretation*—not just the *design*—of risk assessments as a site of contest. A focus on reframing notions of crime and criminal justice has long been at the heart of fights for racial justice. In *Black Feminist Thought*, Patricia Hill Collins writes that "activating epistemologies that criticize prevailing knowledge and that enable us to define our own realities *on our own terms*" is essential to empowering Black women [94]. Prison abolition similarly aims to dismantle carceral discourses and to create alternative, emancipatory ones. Prisons are so ingrained in culture and common sense that "it requires a great feat of the imagination to envision life beyond the prison" [117]. The path toward decarceration therefore requires society "to counter criminological discourses and knowledge production that reify and reproduce carceral logics and practices" [57].

Risk assessments are often hailed in ways that reify and reproduce carceral logics and practices. Yet by expanding the scope of analysis, it is possible to reinterpret risk assessments to demonstrate the limits of dominant anti-discrimination frameworks and to identify a path toward more structural criminal justice reform. The emphasis on substantive equality enables a reform approach that avoids the seemingly intractable bind presented by the impossibility of fairness and the false choice between implementing risk assessments and doing nothing [35, 277, 327]. For when faced with decisions that significantly structure its subjects' lives, the answer is not to optimize the formal fairness of that decision but "to renovate the structure [of the decision] itself, in ways large and small, to open up a broader range of paths that allow people to pursue the activities and goals that add up

to a flourishing life" [165].

There are countless opportunities to renovate the structure of criminal justice decisions and thereby escape the "impossible" choices of risk assessments. Criminal justice institutions can change what interventions are made based on risk assessments, responding to risk with support rather than punishment, as described in Part II. Reducing pretrial detention and mandatory minimums [174] (reforms which polls suggest are popular [41, 168, 175]) can further diminish the harms and scope of risk assessments. The gaze of risk assessments can be turned from defendants to the actors and institutions that comprise the criminal justice system [72, 110], enabling a more structural view of the system's operations. Governments can implement policies that reduce the risk of general, pretrial, and inmate populations [118, 225, 281, 484], thus diminishing the role for punitive responses to risk. The logic behind such reforms is not to reject risk assessments in favor of the status quo, but to reject the structures underlying risk assessments in favor of decarceral and non-punitive structures.

8.5 Discussion: Algorithmic Fairness and Social Change

Despite their widespread support, risk assessments are based in a deficient theory of change: they provide neither objectivity nor meaningful criminal justice reform. Risk assessments bear no guarantee of reducing incarceration—instead, they are more likely to legitimize the criminal justice system's carceral logics and policies. Yet because support for risk assessments emerges in part from the sociotechnical imaginary that sees all problems as solvable with technology, critiques that articulate the technical limits of risk assessments will likely be met by calls for "better" risk assessments. It is therefore necessary to pursue an "epistemic reform" that challenges the *discourses* rather than the *technical specifications* of risk assessments. The impossibility of fairness can be reinterpreted as an incompatibility of equality, demonstrating how mechanisms of formal equality in an unequal society lead to substantive inequality. Seen in this light, risk assessments demonstrate the limits of formalist, colorblind proceduralism and suggest a more expansive and structural approach to criminal justice reform.

These arguments highlight the myopia of "fairness" as a framework for evaluating the social impacts of algorithms. Although researchers have tended to equate technical and social notions of fairness, fairness in its myriad and conflicting meanings cannot be reduced to a single mathematical definition that exists in the abstract, apart from social, political, and historical context [200]. Guaranteeing these technical conceptions of fairness is therefore drastically insufficient to guarantee—or even reliably promote—just social outcomes. Two issues in particular stand out.

First, algorithmic fairness sidelines the social contexts in which decision-making occurs. It treats fairness as a matter of making accurate predictions but does not interrogate the structures behind why certain people are prone to the outcome being predicted or what actions are taken based on predictions. With some exceptions [27, 35, 101, 158], algorithmic fairness debates and metrics hinge on comparing false predictions across groups [17, 35, 226, 278, 327], the implication being that a perfectly accurate model would eliminate the core problem of unfairness. Indeed, recent scholarship asserts that "[t]he most promising way to enhance algorithmic fairness is to improve the accuracy of the algorithm" [226] and that "[t]he largest potential equity gains may come from simply predicting more accurately than humans can" [277].

Although there are fairness benefits to be achieved through improving the accuracy of predictions, the emphasis on accuracy reveals how algorithmic fairness is primarily concerned with Human Bias rather than Population Inequity. Accurate predictions about an unequal society are typically seen as fair. Yet even a "perfect" risk assessment will reinforce the racial discrimination that has structured all aspects of society. As such, algorithmic fairness narrows the scope of judgments about justice, removing structural considerations from view. In this way, algorithmic fairness "mirror[s] some of antidiscrimination discourse's most problematic tendencies," most notably the "fail[ure] to address the very hierarchical logic that produces advantaged and disadvantaged subjects in the first place" [230]. Avoiding the perpetuation of historical harms through algorithms "will often require an explicit commitment to substantive remediation rather than merely procedural remedies" [27].

Second, algorithmic fairness fails to account for the trajectory of social change facilitated by algorithms. Al-

though often intended to improve society, algorithms can—even when satisfying fairness criteria—perpetuate or exacerbate inequities. Evaluations of fairness do not consider the harms of an individualistic approach to reform, the potential of algorithmic decision-making to legitimize unjust systems, or the dangers of conceiving decision-making and reform as technical projects. Instead, an algorithm's fairness is treated as determinative of it having fair social impacts; as long as risk assessments can lead to more accurate or fair decisions, the thinking goes, they are a step in the right direction [35, 191, 474].

Yet creating a more equitable society is not simply a matter of having algorithms generate marginally improved outcomes compared to the status quo—it requires responding to social challenges with holistic responses that promote egalitarian structures and outcomes in both the short and long term [191, 194]. As "an aspirational ethic and a framework of gradual decarceration," abolition aims not to make the criminal justice system more humane while retaining its essential structure, but to reduce the need for (and ultimately eliminate) carceral responses to social disorder [332].

Responsibly developing and evaluating algorithms as tools for social progress requires new methods based in the relationship between technological interventions and social outcomes. First, recognizing the indeterminacy of procedural reforms, reform advocates should avoid deterministic assumptions about the impacts of technology. Rather than viewing technology as a discrete agent of predictable change, reformers should consider the potential for unexpected impacts and should ground any algorithms used within circumstances conducive to reform. For instance, drawing on approaches to limiting legal indeterminacy, the implementation of algorithms could be tied to "sunset provisions" that condition ongoing use of the algorithm to approval based on the results of algorithmic impact assessments [415]. Second, to counter the harms of individualistic decisions and logics, computer scientists must develop new methods that recognize and account for the structural conditions of discrimination, oppression, and inequality. Third, rather than developing tools that are likely to streamline and legitimize existing systems, algorithm developers should thoughtfully consider what interventions will actually be effective at promoting the desired social outcomes. In many cases, typical algorithmic "solutions" may be

counterproductive compared to alternative algorithmic approaches or non-algorithmic reforms.

The challenges *raised* by questions of algorithmic fairness are not—and must not be—limited to the scope of analysis *presented* by algorithmic fairness. Algorithmic decision-making raises fundamental questions about the structure of institutions and the types of reform that are appropriate in response to injustice. Yet as currently constituted, algorithmic fairness narrows these debates to the precise functioning at the decision point itself. This approach overlooks and legitimizes the context that gives structure and meaning to the decision point. In turn, it leads down a path toward dilemmas that, within this scope, appear intractable. Escaping these false choices requires that "we question [our] assumptions and try to look at the issues from another point of view" [341]. Approaching algorithms as sociotechnical imaginaries rather than as discrete technologies enables this expanded scope of analysis. By highlighting the entire context surrounding algorithms as subject to reimagination and reform, this approach avoids the trap of false dilemmas and makes possible more substantive change. Engaging in this manner with today's complex socio-legal-technical environments will inform new paths for algorithms and for reform, in the criminal justice system and beyond.

Chapter 9

Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought

9.1 Introduction

Computer science faces a gap between its desire to do good and the harmful impacts of many of its interventions. The challenge for the field is how to account for social and political concerns in order to more reliably achieve its aims. Our goal in this chapter is to engage with this challenge and to elaborate a positive vision of how computer science can better contribute to society.

To engage in this task, we consider the relationship between "algorithmic thinking" (how algorithms are canonically taught and understood) and "algorithmic interventions" (how algorithms are deployed to address social problems). We are specifically interested in interrogating the influence of "algorithmic thinking" on "algorithmic interventions," and focus on the application of optimization and machine learning algorithms. The divergent meanings of "algorithms" within critical discourse and computer science [139] reflects the differences between

algorithms in theory and the "algorithmic systems—intricate, dynamic arrangements of people and code"—that exist in practice [445]. Understanding the relationship between these two notions of algorithms thus requires approaching algorithms as "'multiples'—unstable objects that are enacted through the varied practices that people use to engage with them" [444].

Algorithmic thinking can be understood as a mode of reasoning: it shapes how computer scientists see the world, understand problems, and develop solutions to those problems. Like all methodologies, algorithmics relies on defining the bounds of its analysis. Considerations that fall within a method's analytic boundaries are the subject of "sharp focus" [441]—but when aspects of the world fall outside these boundaries, a method "has no hope of discovering these truths, since it has no means of representing them" [15]. Thus, as computer science increasingly engages with social and political contexts, the field has come up against the limits of algorithmic thinking: computer science lacks the language and methods to fully recognize, reason about, and evaluate the social aspects and impacts of algorithmic interventions. In turn, even well-intentioned algorithmic interventions are at risk of producing social harms.

Enabling computer science to responsibly navigate its social effects requires several steps: 1) diagnosing the attributes of algorithmic thinking and how those attributes lead to harm, 2) evaluating the potential and limits of current efforts to reform algorithms, 3) describing how the field can expand its epistemic and methodological boundaries, and 4) articulating the tenets of a computer science practice that is evolved based on the concerns raised by affected communities and disciplines such as STS. This chapter takes on each of these tasks in turn.

First, we argue that many of the harms of algorithmic interventions derive from the dominant mode of thinking within computer science, which we characterize as "algorithmic formalism." Algorithmic formalism involves three key orientations: objectivity/neutrality, internalism, and universalism. Although often reasonable (even valuable) within the context of traditional algorithmic work, these orientations can lead to algorithmic interventions that entrench existing social conditions, narrow the range of possible reforms, and impose algorithmic logics at the expense of others. Characterizing these concerns—which draw heavily on STS and critical algorithm

studies—under the banner of formalism provides a path to evaluating and pursuing potential remedies.

Second, we evaluate the dominant approaches to reducing algorithmic harms, such as efforts to promote algorithmic fairness, ethics, and various forms of data and model documentation. Such efforts provide important mechanisms to mitigate certain algorithmic harms. Yet these reforms involve incorporating new processes or metrics into the formal method, and thus do not allow practitioners to transcend formalism itself. Additions of form—most notably, algorithmic fairness—fail to provide the epistemic and methodological tools necessary to fully identify and act upon the social implications of algorithmic work. To solve the chronic failures of algorithmic formalism, computer scientists need new modes of reasoning about the social, both as a terrain of intervention and as an attribute of their own work. This requires an evolution of algorithmic reasoning, expanding the bounds of what it means to "think" algorithmically and "do" algorithmic interventions.

Third, we consider a possible path forward. An epistemic and methodological evolution is a daunting task, and it is not obvious how such a shift could occur or that it would be productive. With this in mind, we draw on our characterization of algorithmic formalism to explore a parallel to formalism in another field—law—and to how legal formalism was addressed with a methodological evolution toward legal realism. From around 1860 through the beginning of the twentieth century, American legal thought was characterized by legal formalism: a project to systematize law around scientific and deductive principles. Because this mode of thought adhered to objective principles but did not consider those principles' actual impacts, its application upheld highly unequal social conditions. These impacts provoked critiques that led to a methodological evolution toward legal realism. Legal realism did not wholly supplant formalism, but instead provided lawyers and judges with additional tools to account for the realities of social life and of the law's impacts. This shift—which expanded the terrain on which law could be evaluated and debated—suggests both a path toward reforming computer science and the utility of such a path.

Fourth, drawing on the lessons of legal realism, we propose a new mode of computer science thinking—"algorithmic realism"—that responds to the concerns raised by STS and related disciplines. Compared to algo-

rithmic formalism, algorithmic realism provides three alternative orientations: a reflexive political consciousness, a porousness that recognizes the complexity and fluidity of the social world, and contextualism. As such, algorithmic realism provides the epistemic and methodological tools to develop algorithmic interventions that question unjust social conditions, expand the range of possible reforms, and account for a wide array of values and goals.

At first glance the law may seem like an unusual place to look for inspiration regarding computer science. With a few exceptions [252, 299], law and computer science are typically seen as in tension, or subject to opposing logics: technology moves "fast" while law is "slow," technology is about "innovation" while law is about "regulation," and so on. Yet several parallels suggest why this comparison is apt. Algorithmic interventions operate in a manner akin to legal ones, often taking the place of (or, more precisely, offering a particular technical form of) legal reforms. Like the law, algorithms are commonly invoked as neutral mechanisms of formalized decision-making. Yet in practice, both are subject to debates regarding the proper role for discretion, ways to combat discrimination, and determinations of the legitimate bases for decision-making. Moreover, the recent surge of enthusiasm for "public interest technology" explicitly follows in the footsteps of (and indeed, takes its name from) a prior movement in legal education [437].

Of course, our goal is not to claim a neat one-to-one correspondence between computer science and law (there certainly are substantial differences), but to point to how the lessons of law can inform computer science. Like computer science, the law involves training in a methodological practice that structures how its practitioners create and evaluate social interventions. Modes of legal thought influence legal interventions in much the same way that modes of algorithmic thought influence algorithmic interventions. Legal scholars have long considered the relationship between the intended and actual impacts of social interventions. Thus, we see the parallel to legal formalism/realism as a way to identify a bridge between the deconstructive critique of algorithmic formalism from STS and a new mode of computer science practice—algorithmic realism—that productively engages with these critiques.

Following the history of law, the distinction between algorithmic formalism and realism does not reflect a rigid dichotomy: the evolution toward realism is an *expansion* of computer science to embrace realist orientations alongside formalist ones, not a wholesale *rejection* of formalism. It is precisely the formalism of algorithmic methods that has enabled many of computer science's most exciting advances [102, 279, 511]. Algorithmic realism provides complementary approaches that make sociotechnical considerations legible and commonplace within computer science thinking. This expanded epistemic and methodological toolkit can help computer scientists to address existing problems more fully and to see new questions.

Nor does the distinction between algorithmic formalism and realism fully characterize the behaviors of computer scientists. In practice, computer scientists are "diverse and ambivalent characters" [444] who blend formalist and realist methods, engaging in "nuanced, contextualized, and reflexive practices" [359] as they "continuously straddle the competing demands of formal abstraction and empirical contingency" [387]. Some computer science subfields (such as CSCW [48]) have long histories of engaging with sociotechnical practices, while others (such as FAT*) are actively developing such methods. We aim to highlight examples of realist-aligned work to help shift such work from exception to standard practice. Nonetheless, computer scientists recognize that the insights of STS and critical algorithm studies fall beyond their own interpretive frames [344]. Even within the FAT* community, critical evaluations of the mathematization of fairness suggest the need for further evolution from formalism towards realism [34, 193, 200, 230, 446].

Of course, a turn toward algorithmic realism would not remedy or prevent every algorithmic harm. Computer scientists are just one set of actors within larger sociotechnical systems that include other people, institutions, policies, and pressures. Algorithmic realism may do little directly to remedy the harms of algorithms deployed through discriminatory public policies, by authoritarian regimes, or under exploitative business models. A great deal of algorithmic work is also done by people without formal computer science training. Algorithmic thinking presents a potent site for reform, however. Computer science plays an influential role in society, both directly through the work of developing algorithmic interventions and indirectly as algorithmic thinking shapes how

scholars (both inside and outside the field), practitioners, and public officials conceive of social challenges and progress [50, 194, 426, 511]. For instance, various public policies and business practices draw on algorithmic reasoning as a way to gain legitimacy [194, 375, 529].

Following STS scholars such as Jasanoff [251] and Winner [512], we aim to trace a middle path between technological determinism and social determinism, exploring the ways in which algorithmic artifacts have politics. We see algorithmic realism not as distinct from sociotechnical systems, but valuable precisely because it situates algorithmic interventions *within* sociotechnical systems. Computer scientists are not the only or the most important actors within sociotechnical systems (nor should they be). Yet reforming such systems requires that computer scientists recognize their positionality and reason about what roles they do (and should) have. Thus, providing computer scientists with the epistemic capacity to navigate the inherent socio-political dimensions of their work is an essential component of sociotechnical reform.

9.2 Algorithmic Formalism

Formalism implies an adherence to prescribed form and rules. The chosen form (e.g., text or numbers) is analyzed according to particular rules, often with the explicit purpose of "constrict[ing] the choice of [a] decisionmaker" or analyst [435]. In literature, for example, formalism involves "the view that the formal properties of a text—its form, structure, grammar, and so forth—define its boundaries" such that "the text stands on its own as a complete entity" [69]. Similarly, formalism in mathematics involves the idea that "mathematics is not a body of propositions representing an abstract sector of reality but is much more akin to a game" where meaning derives from manipulating symbols according to the rules [505].

Formalism is not itself intrinsically bad. It is a method, one that has many virtues. Conceptually, formalizing a problem can lead to analytical clarity. Practically, formalizing a problem can make complex problems tractable. No system that decides how to distribute water, govern educational resources, or predict weather can do so

without in some way formalizing the problem at hand.

Yet formalism also has significant limitations. The danger is not in formal methods *per se*, but in the failure to recognize the limits of formal insights and in the uncritical deployment of formal methods in complex social contexts where their assumptions may be invalid. Because formal knowledge "requires a narrowing of vision," writes James Scott, "the formal order encoded in social-engineering designs inevitably leaves out elements that are essential to their actual functioning" [441]. This narrowing—who or what is left on the epistemic cutting room floor and systemically excluded, or made the focus of overly-simplified analytical scrutiny—involves political decisions about what is and is not important. Formal orders "are often sites of political and social struggles" with "brutal consequences" for those being classified [49].

Formalism is at the core of algorithms. As one canonical algorithms textbook describes, an algorithm is a "well-defined computational procedure" for solving "a well-specified computational problem" [102]. The essential attribute of this reasoning is formalism through abstraction: algorithms require the explicit mathematical articulation of inputs, outputs, and goals. This process of employing abstraction to "formulat[e] a problem to admit a computational solution" is deemed the hallmark of "computational thinking" [511]. As another introductory algorithms textbook explains, "At their most effective, […] algorithmic ideas do not just provide solutions to well-posed problems; they form the language that lets you cleanly express the underlying questions" [279]. Done well, "a clean algorithmic definition can formalize a notion that initially seems too fuzzy and nonintuitive to work with mathematically" [279].

Formalism has been a consistent subject of critique in computing. Brian Smith argued that computer scientists should be attentive to the gulf between the abstractions of models and the complexity of the world [454]. Philip Leith called for computer science to replace computing formalism with a "sociological imagination" [297]. Philip Agre decried "the false precision of formalism" as "an extremely constricting" cognitive style [6]. Pat Langley noted that machine learning research has seen an "increased emphasis on mathematical formalization and a bias against papers that do not include such treatments" [288]. More recently, scholars have highlighted the limits of

abstraction and formalism with regard to algorithmic fairness, articulating the need for a sociotechnical frame [446].

9.2.1 Objective and Neutral

The algorithmic formalist emphasis on objectivity and neutrality occurs on two related levels. First, algorithms are perceived as neutral tools and are often argued for on the grounds that they are capable of making "objective" and "neutral" decisions [194, 257, 395, 245]. Second, computer scientists are seen by themselves and others as neutral actors following the scientific principles of algorithm design from positions of objectivity [191, 240]. Such an ethos has long been prevalent among scientists, for whom objectivity—"the suppression of some aspect of the self, the countering of subjectivity"—has become a widespread set of ethical and normative practices [116].

The orientation of objectivity and neutrality prevents computer scientists from grounding algorithmic interventions in explicit definitions of desirable social outcomes. Social and political[1] concerns are granted little space within algorithmic thinking, leaving data science projects to emerge through the negotiation of technical considerations such as data availability and model accuracy, with "explicit normative considerations [rarely] in mind" [386]. Even efforts that are motivated as contributions to "social good" typically lack a clear explanation of what such "good" entails, instead relying on vague and undefined notions [191]. Computer scientists engaged in algorithmic interventions have argued, "I'm just an engineer" [240] and "Our job isn't to take political stances" [191].

This emphasis on objectivity and neutrality leads to algorithmic interventions that reproduce existing social conditions and policies. For objectivity and neutrality do not mean value-free—they instead mean acquiescence to dominant scientific, social, and political values. Scientific standards of objectivity account for certain kinds of individual subjectivity, but "methods for maximizing objectivism have no way of detecting values, interests,

[1] We invoke politics not in the sense of ideologies, political parties, or elections, but (in a manner akin to Winner [512]) to reference broader debates about "the good"—i.e., the set of processes and dynamics that shape social outcomes and that distribute power among people and groups.

discursive resources, and ways of organizing the production of knowledge" [214]. As such, the supposedly objective scientific "gaze from nowhere" is nothing more than "an illusion" [212]. Neutrality similarly represents an "illusory and ultimately idolatrous goal" that often serves to freeze existing conditions in place [486]. Conceptions of algorithms and computer scientists as objective and neutral launder the perspectives of dominant social groups into perspectivelessness, reinforcing their status as the only ones entitled to legitimate claims of neutrality [**?** **?** 212, 214]. Anything that challenges existing social structures is therefore seen as political, yet reform efforts are no more political than efforts to resist reform or even the choice simply to not act, both of which preserve existing structures.

Predictive policing systems offer a particularly pointed example of how striving to remain neutral entrenches and legitimizes existing political conditions. These algorithms exist on the backdrop of a criminal justice system that is increasingly recognized as defined by racial injustice. Definitions of crime are the products of racist and classist histories that associated Black men with criminality [10, 29, 63, 462]. Moreover, predictive policing is based on the discredited model of "broken windows" policing that has been found to be ineffective and racially discriminatory [63]. In this context, algorithms that uphold common definitions of crime and how to address it are not (indeed, cannot be) removed from politics—they merely *seem* removed from politics. Computer scientists are not responsible for this context, but they are responsible for choosing how they interact with it. When intervening in social contexts steeped in contested histories and politics, in other words, it is impossible for computer scientists to *not* take political stances.

The point is not to argue for a single "correct" conception of the good to which all computer scientists must adhere. It is precisely because a multiplicity of perspectives exists that judgments regarding scientific practices and normative commitments must be explicitly incorporated into algorithm development and evaluation. Yet the epistemic commitments of neutrality and objectivity exclude such considerations from algorithmic reasoning, allowing these judgments to pass without deliberation or scrutiny.

9.2.2 Internalist

Another attribute of algorithmic formalism is internalism: only considerations that are legible within the language of algorithms—e.g., efficiency and accuracy—are recognized as important design and evaluation considerations. The analysis of an algorithm primarily emphasizes its run time (or efficiency), characterizing its behaviors in terms of upper, lower, and tight bounds—all features that can be mathematically defined based on the algorithm's operations [102, 279]. Machine learning algorithms are additionally centered on a corpus of data from which to derive patterns and are evaluated according to accuracy metrics such as area under the curve (AUC). From predictive policing [345] to healthcare [153] to fake news [523], claims regarding an algorithm's effectiveness and quality emphasize accuracy along these metrics. This approach of defining and measuring algorithms by their mathematical characteristics provides little internal capacity to reason about the social considerations (such as laws, policies, and social norms) that are intertwined with these algorithms' performance and impacts.

The internalist emphasis on the mathematical features of algorithms leads to algorithmic interventions based in a technologically determinist theory of social change. Because significant aspects of the social and political world are illegible within algorithmic reasoning, these features are held as fixed constants "outside" of the algorithmic system. In turn, algorithms are proposed as a sole mechanism of social change, with the existing social and political conditions treated as static. For instance, several papers analyzing recidivism prediction tools explicitly describe crime rates in this manner. One describes recidivism prevalence as a "constraint—one that we have no direct control over" [82], while another explains, "the algorithm cannot alter the risk scores themselves" [101]. In an immediate sense it is reasonable to see existing recidivism rates as the backdrop against which a risk assessment makes predictions; yet the recurring practice in computer science of treating these social conditions as a fixed "constraint" exemplifies the internalist assumption that algorithms operate atop a static world.

Internalist reasoning leads to algorithmic interventions that optimize social systems according to existing policies and assumptions, drastically narrowing the range of possible reforms. Algorithmic interventions conceived

through the internalist orientation have a tendency "to optimize the status quo rather than challenge it" [71]. The goal becomes to *predict* (static) distributions of social outcomes in order to make more informed decisions rather than to *shift* (fluid) distributions in order to enable better outcomes. In this vein, an argument made for risk assessments is that "[a]lgorithms permit unprecedented clarity" because they "let us precisely *quantify tradeoffs* among society's different goals" (e.g., fairness and low crime rates); algorithms thereby "force us to make more explicit judgments about underlying principles" [277]. Yet this calculus can be seen to provide "clarity" only if the contingency and contestability of social conditions are beyond consideration. Note the tradeoffs that fall beyond the purview of risk assessments and are therefore rendered irrelevant: for instance, the tradeoff between pretrial detention and due process or the tradeoff between implementing risk assessments and abolishing pretrial detention.

Moreover, because this internalist orientation emphasizes an algorithm's mathematical properties, algorithmic interventions are unable to account for the particular ways that people, institutions, and society will actually interact with algorithms. This is one reason why the deployment of algorithms can generate unintended social outcomes. Algorithmic interventions are thus *indeterminate*: the deployment of an algorithm provides little guarantee that the social impacts expected according to internalist evaluations will be realized. A paradigmatic case involves traffic optimization algorithms, which are modeled on the assumption that increasing road capacity will reduce traffic [368]. Such algorithms have informed urban planning interventions over the past century, from efforts in the 1920s to manage the influx of automobiles [368] to today's visions for self-driving cars [475]. Yet because they rarely account for the second-order effects of their own introduction, these algorithms drastically overestimate the benefits of increasing roadway capacity: in response to more efficient automobile travel, motorists change their behavior to take advantage of the new road capacity, ultimately leading to more driving and congestion [140, 143].

It is impossible for an algorithm to account for every aspect of society or every way that people might respond to it. Every method needs to set boundaries. Yet the choice of where to set those boundaries shapes what factors

are considered or ignored, and in turn shapes the impacts of interventions developed through that method [441]. Internalism enforces a strict frame of analysis, preventing algorithmic interventions from adapting to social considerations that are material to success. Computer scientists therefore need to reason more thoroughly about when certain factors can be ignored and when they must be grappled with.

9.2.3 Universalism

Algorithmic formalism emphasizes an orientation of universalism: a sense that algorithms can be applied to all situations and problems. Popular algorithms textbooks extol the "ubiquitous" applications of algorithms [102] and the "pervasive" reach of algorithmic ideas [279]. An influential computer scientist hails "computational thinking" as "the new literacy of the 21st century," excitedly describing how this mode of thinking "has already influenced the research agenda of all science and engineering disciplines" and can readily be applied in daily life [511]. While some have recognized that there are contexts in which it is better not to design technology [31], the common practice among computer scientists is to focus on *how* to design algorithms rather than *whether* algorithms are actually appropriate in any given context. In fact, when students in data science ethics classes have questioned whether algorithms should be used to address social challenges, they are told that the question is out of scope (J. Geffert, personal communication, April 4, 2019) [516].

This universalist orientation leads to interventions developed under an assumption that algorithms can provide a solution in every situation—an attitude that has been described in recent years as "technological solutionism" [349], "tech goggles" [194], and "technochauvinism" [56]. Algorithmic interventions have been proposed as a solution for problems ranging from police discrimination [161, 194] to misinformation [523, 529] to depression detection [68, 119]. Numerous initiatives strive to develop data science and artificial intelligence for "social good" across a wide range of domains, typically taking for granted with scant justification that algorithms are an effective tool for addressing social problems [191].

Algorithmic interventions pursued under universalism impose a narrow algorithmic frame that structures how

problems are conceived and limits the range of "solutions" deemed viable. Given that "[t]he way in which [a] problem is conceived decides what specific suggestions are entertained and which are dismissed" [128], applying algorithmic thinking to social problems imposes algorithmic logics—namely, accuracy and efficiency—onto these domains at the expense of other values. In "smart cities," for instance, algorithms are being deployed to make many aspects of municipal governance more efficient [194]. Yet efficiency is just one of many values that city governments must promote, and in fact is often in tension with those other values. Inefficient behaviors (such as staff making small talk with residents) can improve a municipality's ability to provide fair social services and garner public trust [522]. More broadly, an emphasis on efficiency in urban life can erode vital civic actions such as deliberation, dissent, and community building [187].

Algorithms can, of course, model a variety of contexts. Efficiency and accuracy are often important factors. But they are typically not the only nor the most important factors. Algorithmic interventions require reasoning about what values to prioritize and what benefits algorithms can provide. However, the universalist orientation prevents computer scientists from recognizing the limits of algorithms and thoroughly evaluating whether algorithms are appropriate.

This uncritical deployment of algorithmic interventions in turn elevates the status of the algorithmic reasoning behind such interventions. Algorithmic formalism has in many ways become the hallmark of what it means to conceive of any problem rigorously, regardless of the many examples of how such thinking faces serious epistemic defects in various social settings. As such, a significant risk of algorithmic formalism is that it contributes to formal methods dominating and crowding out other forms of knowledge and inquiry (particularly local forms of situated knowledge) that may be better equipped to the tasks at hand.

9.3 Formalist Incorporation

One approach to addressing the failures of algorithmic formalism is to incorporate new processes, variables, or metrics into its logic. This process, which we call "formalist incorporation," is particularly appealing to practitioners operating within algorithmic formalism, who tend to respond to critiques of formalizations with calls for alternative formalizations [6]. For example, one paper that describes an algorithmic intervention whose implementation was blocked by community resistance notes, "the legitimate concerns raised by these families can be modeled as objectives within our general formulation and integrated within our framework" [38].

We see many of the recent efforts in the algorithm research and policy communities as examples of formalist incorporation. Specific interventions of this sort include the methods of algorithmic fairness and approaches to improve data and model documentation [177, 233, 343]. Such reforms have significant value and can improve many aspects of algorithms, but they are not designed to provide an alternative mode of reasoning about algorithms. Similarly, although the burgeoning frame of ethics has potential to expand algorithmic reasoning, efforts to promote ethics within computer science and the tech industry have tended to follow a narrow logic of technological determinism and technological solutionism [162, 202, 337, 516]. Because these reforms operate within the logic of algorithmic formalism, they are ultimately insufficient as remedies: formalist incorporation cannot address the failures of formalism itself. When computer scientists raise concerns and engage with social science in this manner, "broader epistemological, ontological, and political questions about data science tools are often sidelined" [344].

We focus here on the methods and research regarding algorithmic fairness, which represents (arguably) the most significant recent change to algorithmic research and practice in response to algorithmic harms. As currently conceived, algorithmic fairness is ill-equipped to address these concerns because it is itself a manifestation of algorithmic formalism via formalist incorporation.

First, algorithmic fairness is grounded in objectivity and neutrality. Fairness is treated as an objective concept,

one that can be articulated and pursued without explicit normative commitments [200]. Approaches to algorithmic fairness often position their goals "in painfully neutral terms" such as "non-discrimination" [230]. Much of the work on fairness points to "bad actors," [230], reinforcing the view that algorithms themselves are neutral. In turn, emphasizing an algorithm's fairness often obscures deeper issues such as unjust practices and policies [194]. What may appear "fair" within a narrow computational scope can reinforce historical discrimination (see Chapter 8).

Second, fairness relies on a narrow, internalist approach: "the mandate within the fair-ML community has been to mathematically define aspects of the fundamentally vague notions of fairness in society in order to incorporate fairness ideals into machine learning" [446]. For example, one paper explicitly "reformulate[s] algorithmic fairness as constrained optimization" [101]. The deployment of an algorithm mathematically deemed "fair" is assumed to increase the fairness of the system in which the algorithm is embedded. For example, predictive policing algorithms and risk assessments have been hailed as remedying the injustices of the criminal justice system [495, 161, 194, 393, 419], with significant energy spent ensuring that these algorithms satisfy mathematical fairness standards. Yet such assessments typically overlook the ways in which these "fair" algorithms can lead to unfair social impacts, whether through biased uses by practitioners (see Chapters 3 and 4), distorting deliberative processes (see Chapter 5), entrenching unjust policies (see Chapter 8), or shifting control of governance toward unaccountable private actors [53, 256, 508].

Third, fairness embodies an attitude of universalism. Attempts to define and operationalize fairness treat the concept universally, with little attention to the normative meaning behind these definitions or to the social and political context of analysis [200, 446]. Much of the algorithmic fairness literature prioritizes portability of definitions and methods across contexts [446]; evaluation tools are designed to fit into any machine learning pipeline [427]. In turn, fairness is applied as the solution wherever algorithmic biases (or other harms) are exposed. For instance, when research exposed that face and gender recognition systems are more accurate on light-skinned men than on dark-skinned women [60], the primary response was to strive for less biased systems

[283, 336, 411, 510, 528], in one case by targeting homeless people of color for facial images [375]. Yet such a pursuit of fair facial recognition does not prevent the systemic harms of this technology—instead, making facial recognition "fair" may legitimize its use under the guise of technical validation [218].

Because of its formalist underpinnings, fair machine learning fails to provide the tools for computer scientists to engage with the critical normative and political considerations at stake when developing and deploying algorithms. Addressing the ways in which algorithms reproduce injustice requires pursuing a new mode of algorithmic thinking that is attentive to the social concerns that fall beyond the bounds of algorithmic formalism.

9.4 Methodological Reform: From Formalism to Realism in the Law

To understand the nature and impacts of an intervention to remedy the limits of formalist reasoning, we turn to the evolution in American legal thought from legal formalism to legal realism.

9.4.1 Legal Formalism

The period from about 1860 through the First World War was one of consensus in American legal thought. The dominant method, called "legal formalism," was the product of concerted effort by legal scholars and judges to promote formal methods in law [268].[2] Legal formalism provided a both a descriptive and a normative account, positing how judicial reasoning does and should occur.

American legal thought in this period was "formal" in several senses. Most fundamentally, law was seen "as a science [that] consists of certain principles or doctrines" [439]. Legal thought aimed to identify, classify, and arrange the principles embodied in legal cases as part of a unified system. Jurists working in this mode tended

[2]The term "legal formalism" was not used by its adherents but was first introduced by legal realists to describe the dominant mode of reasoning they sought to displace. Contemporary legal scholars typically refer to this mode of reasoning (and the period in which it was dominant) as Classical Legal Thought.

to see legal authority as separated along "sharp analytical boundaries—between public and private, between law, politics and morality, and between state and civil society" [269]. Each entity exercised absolute power within its sphere of authority but was not supposed to consider what lay beyond its internalist bounds. Legal formalists favored the application of law along a series of "bright-line" rules; these rigid rules were believed to create a more objective and scientific application of law because they prevented exceptions or context-specific claims. Finally, legal formalism aspired to determinism. It was assumed that a small number of universal principles, derived from natural rights, could be applied to reliably deduce the correct application of law in specific instances [269, 268].

The height of legal formalism coincided with the period of laissez-faire policies and provided reasoning well-suited to defend these policies from progressive challenge. Legal formalism emphasized the autonomy of private citizens and the divide between the authority of the state and that of private actors. From these general principles, judges deduced that efforts to regulate the economy were unconstitutional [490]. In the seminal 1905 case *Lochner v. New York*, the U.S. Supreme Court concluded that a law limiting the working hours of employees represented "unreasonable, unnecessary and arbitrary interference with the right and liberty of the individual to contract" [490]. In his dissent, Justice Oliver Wendell Holmes argued that the Court failed to consider the context of the case, noting, "General propositions do not decide concrete cases" [490]. Legal scholar Roscoe Pound argued that *Lochner* reflected an ignorance of actual working conditions in the United States, which he attributed in part to the "blinders imposed on judges by the 'mechanical' style of judicial reasoning" [398]. Following *Lochner*, it became clear among reform-minded legal scholars that enabling the law to account for the realities of social life necessitated, as a first step, methodological critiques of the formal reasoning that judges used to uphold the status quo.

9.4.2 Legal Realism

The consensus around legal formalism was upended by an alternative mode of thought: "legal realism." Motivated by what they saw as the failure of legal reasoning to account for its real-world impacts, the legal realists

challenged the formalist "jurisprudence of forms, concepts and rules" [269]. They believed that the inability of supposedly well-reasoned legal analysis to address social challenges such as poor working conditions and staggering inequality stemmed from the fact that context-specific realities and the social impact of laws had no place in formal legal analysis.

Achieving social reform therefore required a methodological intervention: a shift in the everyday reasoning of lawyers and judges in order to render social concerns legible in legal thought. Holmes wrote that the "main purpose" of legal realist interventions "is to emphasize certain oft-neglected matters that may aid in the understanding and in the solution of practical, every-day problems of the law" [231]. This pragmatic approach to reform was deeply rooted in the commitment of the legal realists to create a "realistic jurisprudence" focused not on the "paper rules" of black letter doctrine, but the "real rules" that actually described the behavior of courts [306]. Realists aimed to enable the law (and themselves as practitioners of the law) to deal "with things, with people, with tangibles […]—not with words alone" [307]. Rather than simply point out the failures of legal formalism, realist critiques put forward new modes of practical reasoning that overcame the epistemic limitations of formalism and that expanded the commonsense modes of "thinking like a lawyer."

From Universal Principles to Contextual Grounding

A primary legal realist insight was that legal outcomes were not—and could not be—the result of a scientific process. Wesley Hohfeld argued that formal legal thought engaged in deductive error by treating legal principles as universal: "the tendency—and the fallacy—has been to treat the specific problem as if it were far less complex than it really is; and this […] has […] furnished a serious obstacle to the clear understanding, the orderly statement, and the correct solution of legal problems" [231].

The issue arose because efforts to deduce rights and duties from universal principles of liberty or autonomy overlooked how the law was indeterminate (e.g., it could protect "liberty" in multiple competing yet equally plausible ways), confronting decision makers with "a choice which could not be solved by deduction" [269]. In

conventional examples of legal reasoning, legal realists identified instances of "legal pluralism"—the capacity for legal materials (e.g., prior cases, statutes, rules and principles) to render multiple legitimate outcomes due to gaps, conflicts, ambiguities, and circularities within those materials. The resulting indeterminacy and pluralism forced legal actors to make judgments, based on their interpretations and values rather than mechanical procedures, that would structure social dynamics—in effect, making policy.

Realists argued that this adherence to deduction from general principles played a key role in law's complicity with the social harms of the day. Formal legal analysis was evaluated based on whether it correctly identified and applied legal principles. This privileged the formally correct application of principles over the (often unequal) results that such applications created. Realists decried this adherence to "an academic theory of equality in the face of practical conditions of inequality" as methodologically absurd as well as socially harmful [398]. Instead, realists asserted, legal decisions should be evaluated based on their actual impact in their particular context: law should be understood as a means to achieving social ends, not as an end in itself [307]. Unlike philosophy, argued Holmes, the law was not a project of the ideal, but an instrumental means of administering justice in the messy and complex world [234].

From Objective Decisions to Political Assessments

Because cases could not be solved by applying general principles, realists argued that it is impossible to engage in legal decision-making without exercising some degree of subjective judgment. The act of filling gaps in legal reasoning with policymaking was thus infused with politics—the ideological predilections and commitments of the judge. The upshot for realists was not that such expressions of politics are inappropriate, but that they are inevitable.

Realist insights enabled legal practitioners to grapple with the policymaking nature of their work. For example, in cases regarding workers' rights following *Lochner*, judges could no longer reason that judicial interference would be impermissible, because judicial restraint was as much of a political choice as judicial intervention [208].

More broadly, realists displaced the dominance of bright-line rules[3] with a shift towards standards meant to structure reasoning regarding law's social context and impacts.[4] Moving from rules to standards—rendering gaps in deductive legal reasoning more explicit and legible—was one way that the law evolved its methods to incorporate social context and impact into legal doctrines.

From Internalist Boundaries to Porous Analysis

The effort to evaluate the law vis-à-vis its social impact opened up legal analysis to the languages and methods of other disciplines. Legal realists were enthusiastic about filling normative legal gaps with pragmatist philosophy, political science, statistics, sociology, and economics, and decried the failure of law to keep up with developments in "social, economic and philosophical thinking" [399]. They developed limits to legal reasoning *within* legal authority, carving out spaces where law should defer to these disciplines rather than to a judge.

This emphasis on social impact affected legal analysis in two important ways. First, it opened up terrain for the positive program of incorporating "considerations of social advantage" [234] into legal decision-making—to resolve ambiguities in legal materials by looking to social realities. Robert Hale's analysis of industrial workplace conditions typifies this approach [208]. Hale argued that judicial decisions relying on broad commitments to freedom (and opposition to government coercion) to protect "freedom of contract" from workplace unionization flew in the face of Industrial Era workplace conditions, where workers faced extreme coercive pressure from private employers. Hale showed how legal decisions necessarily distribute freedom and coercion among parties, thus necessitating that decisions be made in reference to a broader social objective.

Second, the focus on impact shifted legal thinking toward considering how opinions and laws would play out in practice. Holmes argued that legal inquiry should concern itself with the messy administration of justice among real-world actors. Legal scholars and judges should therefore think of law as would "a bad man" who

[3]E.g., "If you are on the property of another without consent, they are not liable for any injury you may suffer under trespass."

[4]E.g., "Under certain conditions, it may be socially desirable for us to enforce liability even under conditions of trespass: for example, if the harm came to a child lured onto the property by an attractive nuisance."

is not motivated by "the vaguer sanctions of conscience" but only the "material consequences" that may befall him if he runs afoul of the law [234]. To assess whether a law is good or bad, in other words, legal thinkers ought to anticipate the behavior of actors looking to take advantage of the law.

The Realist Evolution of Legal Common Sense

Realist critiques and proposals were controversial and spurred intense debate [269, 307]. Moreover, realist interventions did not provide a silver bullet to the intractable challenge of administering justice through law. Nor did legal realism fully supplant legal formalism: many formalist orientations remain common in American legal thought (and have in recent decades regained prominence in many areas).

Nonetheless, legal realism provided the methodological basis for profound legal reform. Realist methods enabled progressive changes in private law, provided the intellectual foundations for the administrative state, and led to the overturn of *Lochner v. New York* and the subsequent creation of American labor law [268]. Perhaps legal realism's most significant contribution was expanding the epistemic and methodological terrain on which legal reasoning and legal debate could occur. By the 1950s, law students became adept at reasoning about the limitations of law and at making arguments about the policy effects of legal decisions. American legal pedagogy "deeply absorbed the basic idea that the validity of laws should be measured, in part, in terms of their social and economic effects" [342]. Realist methods remain highly influential and have provided the intellectual foundation for several subsequent and ongoing efforts to expand legal thinking, including critical legal studies [267], critical race theory [109], law and economics [396], and law and political economy [204].

9.5 Algorithmic Realism

Recognizing the dangers of algorithmic formalism and the lessons of legal realism, we turn now to articulating the principles of algorithmic realism. These aspirational attributes counter the orientations of algorithmic formalism, with particular attention to preventing (or at least mitigating) the harms it can produce. As the case of legal

thought demonstrates, such a shift can productively enhance a discipline's epistemic and methodological ability to engage with the social. While no mode of reasoning can avoid imposing its logic on the world, self-conscious modes can expand their internal logic to explicitly reason about their effects on the world.

9.5.1 Political

Rather than strive for unattainable notions of objectivity and neutrality, algorithmic realism emphasizes that algorithmic interventions are inherently political. This does not entail computer science entirely abandoning objectivity and its practices, such as the norm against manipulating data in order to generate desired results. Instead, it means interrogating the types of subjectivity that typically fly under the radar of "objective" practice: choices such as formulating research questions, selecting methodologies and evaluation metrics, and interpreting results.

This political orientation enables computer scientists to reflect on the normative commitments and outcomes of algorithmic interventions. Rather than creating paralysis, with computer scientists unsure how to be neutral and objective when doing so is impossible, algorithmic realism provides a language to reason about political commitments and impacts as part of what it means to "do" algorithms. First, freed from the strict imperative to be neutral and objective, computer scientists can interrogate the ways in which their assumptions and values influence algorithm design. This reflexive turn can help computer scientists—regardless of their particular normative commitments—better reason about the relationship between their design choices, their professional role, and their vision of the good. Such reflection should occur through open discussion and deliberation, forming a central component of the research process. Second, algorithmic realism shifts the primary focus of algorithmic interventions from the quality of an algorithmic system (in an internalist sense) to the social outcomes that the intervention produces in practice. No matter how technically advanced or impressive a system is, its success under an algorithmic realist frame is defined by whether that system actually leads to the desired social changes.

This approach enables interventions that question rather than uphold unjust social conditions and policies.

Several approaches can inform such development of algorithms. The schema of "reformist" and "non-reformist" reforms, articulated by social philosopher André Gorz, provides a way to evaluate interventions based on their political implications [189]. While a reformist reform "subordinates its objectives to the criteria of rationality and practicability of a given system and policy," a non-reformist reform "is conceived not in terms of what is possible within the framework of a given system and administration, but in view of what should be made possible in terms of human needs and demands." Designers Anthony Dunne and Fiona Raby classify design into two categories: affirmative design, which "reinforces how things are now," and critical design, which "rejects how things are now as being the only possibility" [142]. A related framework is "anti-oppressive design," which orients "the choice of a research topic, the focus of a new social enterprise, or the selection of clients and projects" around challenging oppression [457]. Similarly, the Design Justice Network provides ten design principles that include "prioritize design's impact on the community over the intentions of the designer" [360].

These frameworks show that recognizing algorithmic interventions as political does not prevent computer scientists from doing computer science—instead, doing so can help them incorporate normative reflection into the methods and questions that drive their work. With this in mind, computer scientists can ask a variety of questions to inform their practice: Would the implementation of this algorithm represent a reformist or non-reformist reform? Is the design of this algorithm affirmative or critical? Would providing our project partner with this algorithm entrench or challenge oppression? Is the project prioritizing outcomes over my intentions? Will this algorithm empower the communities it affects?

An example of the expanded practical reasoning that a political orientation provides involves burgeoning activism among employees of technology companies against developing algorithmic interventions for use by the United States Departments of Defense and Homeland Security [179]. Rather than perceiving themselves as "just an engineer" [240], these computer scientists recognize their position within larger sociotechnical systems, perceive the connection between developing an algorithmic intervention and the political and social outcomes of those interventions, and hold themselves (and their companies) accountable to the impacts of the algorithms

they develop. Building on this movement, in 2019, thousands of computer science students from more than a dozen U.S. universities pledged that they would not work for Palantir due to its partnerships with Immigration and Customs Enforcement (ICE) [338].

9.5.2 Porous

Recognizing algorithmic formalism's limited ability to characterize sociotechnical systems, algorithmic realism is porous, expanding the range of considerations deemed relevant to algorithm design and evaluation. Factors that were previously beyond the internalist algorithmic frame become central to what it means to have knowledge or make claims about algorithms. A porous approach to algorithms means that formalist considerations (e.g., accuracy, efficiency, and fairness) are recognized as necessary but *no longer sufficient* to define the efficacy or quality of an algorithm—additional modes of analysis are essential. As in law, realism entails both an appreciation of the insights of other fields and a willingness, where appropriate, to carve out spaces of deference to those fields.

This porous orientation allows for algorithmic interventions that eschew technological determinism and instead recognize the contingency and fluidity of the social world. It makes legible the potential for social and policy change in addition to (or instead of) technological change. This does not mean adopting a mantra of social determinism, believing that social systems will evolve irrespective of technology. Instead, a porous approach to algorithmic interventions follows an STS understanding of how "the realities of human experience emerge as the joint achievements of scientific, technical and social enterprise" [251].

This porous orientation gives computer scientists the capacity to widen rather than narrow the range of possible reforms. Rather than optimizing existing systems under the assumption of a static society, computer scientists can develop interventions under the recognition of a fluid society. Several projects exemplify this approach. For example, instead of developing predictive policing or risk assessment algorithms that treat risk levels and policy responses as static, computer scientists have developed algorithms to reduce the risk of crime and violence through targeted and non-punitive social services (see Chapter 7). In other contexts, computer

scientists have subordinated their priorities to broader communities, helping to empower groups advocating for change [20, 104, 129, 265, 335, 324, 436].

Furthermore, by bringing the social world into the algorithmic frame, a porous orientation allows for algorithmic interventions that recognize and account for indeterminacy. Under algorithmic realism, "good" algorithm design means not simply designing to promote desired outcomes, but defining what outcomes are desirable and undesirable, understanding how potential harms could arise, and developing anticipatory mechanisms to prevent or mitigate those outcomes. By incorporating these considerations as essential to algorithm design, algorithmic realism casts practices such as failing to consider how users interact with an algorithm as no less negligent than failing to test a model's accuracy.

Although it is impossible to fully account for indeterminacy or to guarantee that an intervention will have particular impacts, scholarship from STS and critical algorithm studies provides valuable starting points for analyzing the relationship between algorithmic interventions and social impacts. The Social Construction of Technology (SCOT), for example, argues that new technologies contain numerous potential interpretations and purposes; how a technology stabilizes (in "closure") depends on the social groups involved in defining that technology and the relative resources each has to promote its particular vision [392]. Co-production more richly articulates the intertwined nature of technology and social conditions, noting identities, institutions, discourses, and representations as particularly salient pathways of social and technological change [251]. A great deal of other recent work has documented the particular ways in which the design, application, and use of algorithms can exacerbate marginalization and inequality [56, 154, 194, 216, 365, 372].

Taking these approaches as a guide, numerous questions can inform computer scientists' understanding of how an algorithm will interact with and impact communities in a given context. These include: Who are the relevant social actors? What are their interests and relative amounts of power? Which people need to approve this algorithm? What are their goals? On whose use of the algorithmic system does success depend? What are their interests and capabilities? How might this algorithm affect existing scientific, social, and political discourses

or introduce new discourses?

This approach has particular value in anticipating and preventing harmful social impacts of algorithms. Just as Holmes urged legal scholars and judges to evaluate laws in light of how they will be carried out in practice, so too should computer scientists evaluate algorithmic interventions through the lens of how people may actually apply them. For example, recognizing how police use of algorithms can distort interventions toward surveillance and punishment, some researchers developing algorithms to identify people at risk of involvement in crime or violence explicitly articulate their commitment to partnering with community groups and social service providers rather than with law enforcement [30, 171].

9.5.3 Contextual

In contrast to the universalism of algorithmic formalism, algorithmic realism is grounded in contextualism, emphasizing the need to understand social contexts in order to determine the validity of any algorithmic intervention. Rather than question *how* a situation can be modeled and acted upon algorithmically, a contextual approach questions *to what extent* a situation can be modeled and should be acted upon algorithmically. Context is defined here not in a positivist sense of data that can be incorporated into algorithms, but in a broader sense entailing the social relations, activities, and histories that shape any particular setting [138]. Gleaning context therefore requires a porous approach rather than an internalist focus on data [50, 138, 443, 171]. Such context is essential to understanding relationships and behaviors in sociotechnical systems [323, 363].

A contextual orientation allows computer scientists to avoid solutionism and instead take an agnostic approach to algorithmic interventions. Agnosticism entails approaching algorithms instrumentally, recognizing them as just one type of intervention, one that cannot provide the solution to every problem. In other words, an agnostic approach prioritizes the social impacts of reform, regardless of the role played by algorithms—it is agnostic as to the means, but not the ends. This approach can help not just to avoid harmful algorithms, but also to place algorithms alongside institutional and policy reforms in order to robustly promote well-articulated social ends.

For even in contexts where algorithms can help to address social challenges, they cannot do so in isolation: the most impactful algorithmic interventions occur when algorithms are deployed in conjunction with policy and governance reforms [194].

This approach also allows algorithmic thinking to be incorporated into social and policy reform efforts without requiring the deployment of an algorithm and the imposition of algorithmic logics. Contextualism makes legible questions about whether algorithms can capture the essential aspects of a real-world context and whether algorithms can generate the desired social impacts. Computer scientists pursuing interventions through a contextual approach can pose numerous questions: What elements of this context does an algorithmic approach capture and overlook? What values are important for any solution? To what extent can an algorithm account for those values? How does an algorithm compare to other reforms in terms of producing better outcomes? If the answers to these questions suggest a significant divide between the context and an algorithm's ability to model and improve that context, then it is likely that an algorithmic intervention is an ill-advised approach to providing the desired social benefits.

To see this in practice, consider the experience of the author while working as a data scientist with a municipal Emergency Medical Services (EMS) department. The author was asked to improve ambulance response times with data analytics. The instinct of an algorithmic formalist, following a universalist orientation, would be to develop an algorithm that optimizes ambulance dispatch [47, 244, 248, 525]. Yet when the author studied the context of the problem, it became clear that such a "solution" would not fit into EMS's operations nor would it address the underlying issues generating long response times. The author's analysis revealed that significant resources were being deployed to 911 calls for people struggling with homelessness, mental illness, and drug addiction. These individuals did not require the acute medical care that EMS was providing (at the expense of providing it for other incidents); instead they needed social services that EMS was ill-equipped to provide. It became clear that ambulance response efficiency was a limited frame for understanding (and thus reforming) EMS's operations: the efficiency of ambulance responses said nothing about the broader goal of providing

services that address people's needs.

Although a dispatch optimization algorithm may perform well along formalist metrics of efficiency, such an algorithm would have failed to address the underlying issue. The author instead worked with EMS to create a new unit of EMTs who would respond to these incidents via bicycle or car and be specially trained to connect people to local social services; the parameters of when and where this unit would operate were determined by analyzing EMS incident data. Notably, the ultimate intervention was *not* to integrate an algorithm into existing procedures: a policy change informed by data was better suited to improve both efficiency and service quality. Rather than representing a failure to take advantage of algorithms, this effort was recognized as a positive collaboration that integrated data analysis and institutional context to improve social services.

9.6 Discussion

The numerous and significant harms of algorithms may appear to be the result of computer scientists failing to follow best practices. Yet our articulation of algorithmic formalism describes how these outcomes are due to the logic of algorithmic thinking itself, not an imperfect or malevolent application thereof. The chronic tunnel vision of algorithmic formalism can lead to harmful outcomes despite good intentions and following current best practices. Remedying these failings requires not incorporating new variables or metrics (such as fairness) into the formal method but instead introducing new epistemic and methodological tools that expand the bounds of what it means to "do" algorithms.

Algorithmic realism represents this evolution in algorithmic thought, providing new modes of practical reasoning about the relationship between algorithms and the social world. The realist orientations described here provide important starting points for computer scientists and others pursuing algorithmic interventions. Following the political orientation, practitioners should consider what assumptions and values they may be taking for granted and what normative commitments they want their intervention to embody. Following the porous

orientation, practitioners should consider what theory of change motivates their work and how to responsibly account for unexpected impacts. Following the contextual orientation, practitioners should consider what goals are central to a given context and whether an algorithm actually provides an appropriate intervention. In a realist mode of reasoning, all of these questions are seen as integral to rigorous algorithmic work rather than as beyond the scope of algorithmic design. These realist practices will enable the field not just to avoid harmful impacts, but also to identify new research questions and directions to pursue.

As in law, algorithmic realism is not meant to provide a wholesale rejection of formal methods nor will it provide a wholesale solution to the intractable challenges of designing just algorithmic systems. Even to the extent that the turn to algorithmic realism is motivated by a broader program of social reform (à la the turn to legal realism), new epistemic and methodological tools cannot by themselves achieve a vision of the good, let alone determine which vision of the good to work towards. Nonetheless, algorithmic realism can help computer scientists reflexively approach their work in light of their larger normative commitments and the impacts of algorithmic systems. As such, algorithmic realism enables computer scientists to reason well about doing good.

Chapter 10

Conclusion

As algorithms become commonplace across a wide range of policy domains, it is necessary to expand the criteria incorporated into the design and implementation of these systems. A methodology of drawing on legal and social theory is a necessary step toward achieving more responsible and effective approaches to algorithms and social change. Looking to fields that have long engaged with the relationship between social interventions and social impacts can therefore inform new ways of understanding, developing, and governing algorithms.

One area of work will be to explore how algorithmic interventions can be recognized as embodying legal strategies for social change. This requires reconceptualizing the relationship between algorithms and law: neither wholly distinct nor wholly interchangeable, the two instead represent related approaches to rule-based decision making and managing discretion. Although law and technology are typically seen as being in tension, recognizing the ways in which law and technology resemble similar mechanisms of social ordering suggests the potential for legal theory and STS to inform one another [252]. Algorithmic interventions often represent a particular form of what have in the past have been policy reforms. And as with algorithms, the law is subject to critiques that expose the limits of its supposed objectivity and neutrality.

A second area of work will be to consider how the analogy between algorithmic and legal interventions can inform our understanding and governance of algorithms. This will involve characterizing what types of social impacts algorithmic interventions are capable of generating and how those impacts relate to the design and governance of algorithms. By analyzing algorithmic interventions through the lens of STS and law, it may be possible to identify what types of social change these interventions have produced and to evaluate the potential and limits of such reforms.

Doing so can provide several important insights for studying and governing algorithms. First, it can indicate when and where algorithms present an appropriate reform strategy. Despite the adoption of algorithms across a wide range of policy domains, algorithms are more clearly suited toward certain types of problems than others. Responsible deployment of algorithms requires understanding where algorithms are appropriate and where they are not. One recent paper suggests four particular roles for algorithms [1]. Second, this analysis can reveal more systemic issues that are often glossed over by algorithmic interventions. The deployment and impacts of algorithms are often grounded in broader issues (such as austerity and a lack of political will for alternative reforms) in addition to the technical capacities and affordances of algorithms. Analyzing the delta between contexts in which algorithms are appropriate and are deployed may therefore highlight where structural issues are being overlooked or obfuscated. Third, this analysis can inform governance strategies for algorithms. Recognizing algorithms as akin to prior legal forms can point the way toward adapting existing governance models from other domains to algorithms. For instance, drawing on approaches to limiting legal indeterminacy, the implementation of algorithms could be tied to "sunset provisions" that condition ongoing use to approval based on the results of algorithmic impact assessments.

The foundations of these first two areas will enable algorithmic interventions that are better equipped to improve society. This requires developing a sociotechnical computer science practice that thinks rigorously and reflexively about the role of algorithmic interventions in producing social change—in other words, that is as rigorous regarding social context and impacts as it is regarding algorithms themselves.

One area of work involves developing methods to improve human-algorithm collaboration in decision-making. As I have shown in this thesis, there exist significant breakdowns in decision making when predictive models are presented to humans. Achieving just decisions will require adapting existing theories of just decision making to the context of algorithm-informed decisions as well as following a sociotechnical research practice that starts with studying what actually improves human decision making and working backwards from there. For instance, rather than presenting people with predictive models that duplicate their decisions, it may be better to present people with algorithmic tools that provide context and feedback. It is also important to consider the many different types of decisions in which human-algorithm collaborations arise and how the effectiveness of different approaches may vary across those tasks.

This approach also suggests future research directions with regards to to the methods described in Chapter 7. First, rather than just identifying high-risk individuals, can we identify nodes or edges in the network where intervention could have the largest spillover effects? It is not necessarily the case that the highest-risk nodes are also the places where intervention will have the largest net effects. It could be that it is more effective to intervene (i.e., provide resources to reduce the likelihood of violence) not on the highest risk individuals, but on the individuals most likely to "spread" violence through a social network. This requires also attaining a greater understanding of the spillover effects of different types of interventions [515]. Second, is it possible to alter the dynamics of contagion over the network? Given dynamics of violence transmission such as those described in Chapter 7, what would it look like to change these dynamics or change the structure of the network? In other words, rather than designing interventions at the level of individuals, it could be more effective to design broader, population-level interventions that reduce the dynamics of transmission or make the network less conducive to transmission.

Another topic that deserves particular attention is the extent to which algorithms can account for structural inequality when making decisions. Additional scholarship is necessary to understand the relationship between accuracy and fairness, and in particular the potential limits and harms of accuracy when operating in settings

with existing inequities. Drawing on debates about different legal strategies for addressing discrimination will inform efforts to develop algorithmic interventions that remediate rather than reproduce historical inequities. Considering the limits of accuracy is also important for evaluating algorithm-in-the-loop systems: accuracy is not the only value in decision making, so positioning accuracy as the sole or most important criterion may prevent us from fully capturing the quality of human-algorithm collaborations. An important open question is how to design experiments that study algorithm-in-the-loop decision making without inappropriately privileging easily quantifiable outcomes (such as accuracy and efficiency).

Finally, it will be essential to explore methods for developing and evaluating algorithms through democratic deliberation. The increasing centrality of algorithms in public policy make the development and evaluation of these algorithms an important site for democratic decision making. Yet there do not yet exist robust models for how to have effective public deliberations about digital technology; for instance, programs such as open data have not created the rich civic engagement often aspired to. The rich literature on both theories of democracy and empirical studies of civic engagement efforts will be an important starting point for enhancing civic participation and control regarding algorithms. Existing mechanisms for participatory design and decision making—such as charrettes and participatory budgeting—present promising models for incorporating democratic input into algorithmic interventions.

References

[1] Abebe, R., Barocas, S., Kleinberg, J., Levy, K., Raghavan, M., and Robinson, D. G. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 252–260.

[2] Abrams, D. S., Bertrand, M., and Mullainathan, S. Do Judges Vary in Their Treatment of Race? *The Journal of Legal Studies 41*, 2 (2012), 347–383.

[3] Adamic, L. A. The Small World Web. In *Research and Advanced Technology for Digital Libraries* (1999), S. Abiteboul and A.-M. Vercoustre, Eds., Springer Berlin Heidelberg, pp. 443–452.

[4] Adams, J., Moody, J., and Morris, M. Sex, drugs, and race: How behaviors differentially contribute to the sexually transmitted infection risk network structure. *American Journal of Public Health 103*, 2 (2013), 322–329.

[5] Agan, A., and Starr, S. Ban the Box, Criminal Records, and Racial Discrimination: A Field Experiment. *The Quarterly Journal of Economics 133*, 1 (2017), 191–235.

[6] Agre, P. E. Toward a Critical Technical Practice: Lessons Learned in Trying to Reform AI. In *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*, G. C. Bowker, S. L. Star, W. Turner, and L. Gasser, Eds. 1997.

[7] Albert, R., and Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics 74*, 1 (2002), 47.

[8] Albright, A. If You Give a Judge a Risk Score: Evidence from Kentucky Bail Decisions. *The John M. Olin Center for Law, Economics, and Business Fellows' Discussion Paper Series 85* (2019).

[9] Albury, R. *The Politics of Objectivity*. Deakin University Press, 1983.

[10] Alexander, M. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press, 2012.

[11] Alkhatib, A., and Bernstein, M. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2019), CHI '19, ACM, pp. 530:1–530:13.

[12] American Bar Association. *ABA Standards for Criminal Justice: Pretrial Release*, 3 ed. 2007.

[13] Ames, M. G. *The Charisma Machine: The Life, Death, and Legacy of One Laptop Per Child.* MIT Press, 2019.

[14] Anagnostopoulos, A., Kumar, R., and Mahdian, M. Influence and Correlation in Social Networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008).

[15] Anderson, E. Toward a Non-Ideal, Relational Methodology for Political Philosophy: Comments on Schwartzman's "Challenging Liberalism". *Hypatia 24*, 4 (2009), 130–145.

[16] Angwin, J., and Larson, J. Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say. *ProPublica* (2016).

[17] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias. *ProPublica* (2016).

[18] Aral, S., Muchnika, L., and Sundararajana, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences 106*, 51 (2009), 21544–21549.

[19] Arrieta-Kenna, R. 'Abolish Prisons' Is the New 'Abolish ICE'. *Politico* (2018).

[20] Asad, M. Prefigurative Design As a Method for Research Justice. *Proceedings of the ACM on Human-Computer Interaction. 3*, CSCW (2019), 200:1–200:18.

[21] Atkinson, C. Do Not Resist. *Vanish Films* (2016).

[22] Austin, A. The Presumption for Detention Statute's Relationship to Release Rates. *Federal Probation 81* (2017), 52.

[23] Austin, J. Evaluation of Broward County Jail Population: Current Trends and Recommended Options.

[24] Bakshy, E., Messing, S., and Adamic, L. Exposure to ideologically diverse news and opinion on Facebook. *Science* (2015), 1130–1132.

[25] Baldus, D. C., Pulaski, C., and Woodworth, G. Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience. *Journal of Criminal Law and Criminology 74*, 3 (1983), 661–753.

[26] Baradaran, S. Restoring the Presumption of Innocence. *Ohio State Law Journal 72* (2011), 723–776.

[27] Barocas, S., and Selbst, A. D. Big Data's Disparate Impact. *California Law Review 104* (2016), 671–732.

[28] Barry-Jester, A. M., Casselman, B., and Goldstein, D. The New Science of Sentencing. *The Marshall Project* (2015).

[29] Baum, D. Legalize It All. *Harper's Magazine* (2016).

[30] Bauman, M. J., Boxer, K. S., Lin, T.-Y., Salomon, E., Naveed, H., Haynes, L., Walsh, J., Helsby, J., Yoder, S., and Sullivan, R. Reducing Incarceration through Prioritized Interventions. In *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies* (2018), COMPASS '18, ACM, pp. 6:1–6:8.

[31] Baumer, E. P., and Silberman, M. S. When the implication is not to design (technology). In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2011), CHI '11, ACM, pp. 2271–2274.

[32] Bayley, D. H. *Police for the Future*. Oxford University Press, 1996.

[33] Bearman, P. S., Moody, J., and Stovel, K. Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology 110*, 1 (2004), 44–91.

[34] Benthall, S. Critical reflections on FAT* 2018: a historical idealist perspective. *DATACTIVE* (2018).

[35] Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* (2018), 1–42.

[36] Berk, R. A. An introduction to sample selection bias in sociological data. *American Sociological Review* (1983), 386–398.

[37] Bertrand, M., and Mullainathan, S. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review 94*, 4 (2004), 991–1013.

[38] Bertsimas, D., Delarue, A., and Martin, S. Optimizing schools' start time and bus routes. *Proceedings of the National Academy of Sciences 116*, 13 (2019), 5943–5948.

[39] Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 377:1–377:14.

[40] Bliss, L. Former Uber Backup Driver: 'We Saw This Coming'. *CityLab* (2018).

[41] Blizzard, R. Key Findings from a National Survey of 800 Registered Voters January 11-14, 2018.

[42] Block, R., and Block, C. R. *Street gang crime in Chicago*. Roxbury, Thousand Oaks, CA, 1995.

[43] Bonczar, T. P. Prevalence of Imprisonment in the U.S. Population, 1974-2001. *Bureau of Justice Statistics Special Report* (2003).

[44] Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., and Fowler, J. H. A 61-million-person experiment in social influence and political mobilization. *Nature 489*, 7415 (2012), 295–298.

[45] Bond-Graham, D., and Winston, A. All Tomorrow's Crimes: The Future of Policing Looks a Lot Like Good Branding. *SF Weekly* (2013).

[46] Bonta, J., and Andrews, D. A. Risk-Need-Responsivity Model for Offender Assessment and Rehabilitation. *Public Safety Canada* (2007).

[47] Boutilier, J. J., and Chan, T. C. Ambulance Emergency Response Optimization in Developing Countries. *arXiv preprint arXiv:1801.05402* (2018).

[48] Bowker, G., Star, S. L., Gasser, L., and Turner, W. *Social Science, Technical Systems, and Cooperative Work: Beyond the Great Divide*. Routledge, 2014.

[49] Bowker, G. C., and Star, S. L. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 2000.

[50] boyd, d., and Crawford, K. Critical Questions for Big Data. *Information, Communication & Society 15*, 5 (2012), 662–679.

[51] Braga, A. A. Serious youth gun offenders and the epidemic of youth violence in Boston. *Journal of Quantitative Criminology 19* (2003), 33–54.

[52] Braga, A. A., and Weisburd, D. L. The effects of focused deterrence strategies on crime: A systematic review and meta-analysis of the empirical evidence. *Journal of Research in Crime and Delinquency 49*, 3 (2011), 323–358.

[53] Brauneis, R., and Goodman, E. P. Algorithmic Transparency for the Smart City. *The Yale Journal of Law & Technology 20* (2018), 103–176.

[54] Brayne, S. Big Data Surveillance: The Case of Policing. *American Sociological Review 82*, 5 (2017), 977–1008.

[55] Brayne, S., and Christin, A. Technologies of Crime Prediction: The Reception of Algorithms in Policing and Criminal Courts. *Social Problems* (2020).

[56] Broussard, M. *Artificial Unintelligence: How Computers Misunderstand the World*. MIT Press, 2018.

[57] Brown, M., and Schept, J. New abolition, criminology and a critical carceral studies. *Punishment & Society 19*, 4 (2017), 440–462.

[58] Brustein, J. This Guy Trains Computers to Find Future Criminals. *Bloomberg* (2016).

[59] Buhrmester, M., Kwang, T., and Gosling, S. D. Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data? *Perspectives on Psychological Science 6*, 1 (2011), 3–5.

[60] Buolamwini, J., and Gebru, T. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* (2018), A. F. Sorelle and W. Christo, Eds., vol. 81, PMLR, pp. 77–91.

[61] Butler, P. *Let's Get Free: A Hip-Hop Theory of Justice*. The New Press, 2010.

[62] Butler, P. The System Is Working the Way It Is Supposed to: The Limits of Criminal Justice Reform. *The Georgetown Law Journal 104* (2016).

[63] Butler, P. *Chokehold: Policing Black Men*. The New Press, 2017.

[64] Butler, P. Equal Protection and White Supremacy. *Northwestern University Law Review 112*, 6 (2017), 1457–1464.

[65] Butler, P. D. Poor People Lose: Gideon and the Critique of Rights. *Yale Law Journal 122* (2012), 2176–2204.

[66] Butts, J. A., Roman, C. G., Bostwick, L., and Porter, J. R. Cure Violence: A Public Health Model to Reduce Gun Violence. *Annual Review of Public Health 36*, 1 (2015), 39–53.

[67] Bürkner, P.-C. Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal 10*, 1 (2018), 395–411.

[68] Cacheda, F., Fernandez, D., Novoa, F. J., and Carneiro, V. Early Detection of Depression: Social Network Analysis and Random Forest Techniques. *Journal of Medical Internet Research 21*, 6 (2019), e12554.

[69] Cain, M. A. Problematizing Formalism: A Double-Cross of Genre Boundaries. *College Composition and Communication 51*, 1 (1999), 89–95.

[70] Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A Probabilistic Programming Language. *2017 76*, 1 (2017), 32.

[71] Carr, N. The Limits of Social Engineering. *MIT Technology Review* (2014).

[72] Carton, S., Helsby, J., Joseph, K., Mahmud, A., Park, Y., Walsh, J., Cody, C., Patterson, C. E., Haynes, L., and Ghani, R. Identifying Police Officers at Risk of Adverse Events. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), KDD '16, ACM, pp. 67–76.

[73] Casey, P. M., Warren, R. K., and Elek, J. K. *Using Offender Risk and Needs Assessment Information at Sentencing: Guidance for Courts from a National Working Group*. National Center for State Courts, 2011.

[74] Cavallo, M., and Demiralp, c. A Visual Interaction Framework for Dimensionality Reduction Based Data Exploration. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), CHI '18, ACM, pp. 635:1–635:13.

[75] Centola, D. The Spread of Behavior in an Online Social Network Experiment. *Science 329*, 5996 (2010), 1194–1197.

[76] Chandler, D., Levitt, S. D., and List, J. A. Predicting and preventing shootings among at-risk youth. *The American Economic Review 101*, 3 (2011), 288–292.

[77] Chanenson, S. L., and Hyatt, J. M. The Use of Risk Assessment at Sentencing: Implications for Research and Policy. *Bureau of Justice Assistance* (2016).

[78] Chayes, A., Fisher, W., Horwitz, M., Michelman, F., Minow, M., Nesson, C., and Rakoff, T. Critical Perspectives on Rights.

[79] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., Li, M., Xie, J., Lin, M., Geng, Y., and Li, Y. xgboost: Extreme Gradient Boosting.

[80] Cheng, J., Adamic, L., Dow, P. A., Kleinberg, J. M., and Leskovec, J. Can cascades be predicted? In *Proceedings of the 23rd International Conference on World Wide Web* (2567997, 2014), ACM, pp. 925–936.

[81] Chong, V. E., Smith, R., Garcia, A., Lee, W. S., Ashley, L., Marks, A., Liu, T. H., and Victorino, G. P. Hospital-centered violence intervention programs: a cost-effectiveness analysis. *The American Journal of Surgery 209*, 4 (2015), 597–603.

[82] Chouldechova, A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data 5*, 2 (2017), 153–163.

[83] Christakis, N. A., and Fowler, J. H. The spread of obesity in a large social network over 32 years. *New England Journal of Medicine 357* (2007), 370–379.

[84] Christin, A. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data & Society 4*, 2 (2017), 2053951717718855.

[85] Christoffel, K. K. Firearm injuries: Epidemic then, endemic now. *American Journal of Public Health 97*, 4 (2007), 626–629.

[86] Ciccolini, J., and Conti-Cook, C. Rationing Justice: Risk Assessment Instruments in the American Criminal Justice System. *EuropeNow* (2018).

[87] Cicurel, R. Motion to Exclude Results of the Violence Risk Assessment and All Related Testimony and/or Allocution Under FRE 702 and Daubert v. Merrell Dow Pharmaceuticals.

[88] Citron, D. K. Technological Due Process. *Washington University Law Review 85* (2007), 1249.

[89] Clark, E., Ross, A. S., Tan, C., Ji, Y., and Smith, N. A. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (New York, NY, USA, 2018), IUI '18, ACM, pp. 329–340.

[90] Cohen, J., Cork, D., Engberg, J., and Tita, G. The role of drug markets and gangs in local homicide rates. *Homicide Studies 2*, 3 (1998), 241–262.

[91] Cohen, J., and Tita, G. Diffusion in homicide: Exploring a general method for detecting spatial diffusion processes. *Journal of Quantitative Criminology 15*, 4 (1999), 451–493.

[92] Cohen, T. H., Pendergast, B., and VanBenschoten, S. W. Examining overrides of risk classifications for offenders on federal supervision. *Federal Probation 80*, 1 (2016), 12.

[93] Cohen-Cole, E., and Fletcher, J. M. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ, 337* (2008).

[94] Collins, P. H. *Black Feminist Thought: Knowledge, Consciousness, and the Politics of Empowerment*. Routledge, 2000.

[95] Cook, P. J., and Laub, J. H. After the epidemic: Recent trends in youth violence in the United States. *Crime and Justice v29* (2002), 1–37.

[96] Cooper, A., and Smith, E. L. Homicide Trends in the United States, 1980-2008. *U.S. Department of Justice, Bureau of Justice Statistics* (2011).

[97] Coppock, A. Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach. *Political Science Research and Methods 7*, 3 (2019), 613–628.

[98] Corbett-Davies, S., and Goel, S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).

[99] Corbett-Davies, S., Goel, S., and González-Bailón, S. Even Imperfect Algorithms Can Improve the Criminal Justice System. *The New York Times* (2017).

[100] Corbett-Davies, S., Pierson, E., Feller, A., and Goel, S. A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. *The Washington Post* (2016).

[101] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017), ACM, pp. 797–806.

[102] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. *Introduction to Algorithms*. MIT Press, 2009.

[103] Corsaro, N., and Engel, R. S. Most challenging of contexts: Assessing the impact of focused deterrence on serious violence in New Orleans. *Criminology & Public Policy 14* (2015), 471–505.

[104] Costanza-Chock, S., Wagoner, M., Taye, B., Rivas, C., Schweidler, C., Bullen, G., and Project, t. T. #MoreThanCode: Practitioners reimagine the landscape of technology for justice and equity.

[105] Cover, R. M. Violence and the Word. *Yale Law Journal 95* (1986), 1601–1629.

[106] Covert, B. America Is Waking Up to the Injustice of Cash Bail. *The Nation* (2017).

[107] Cowgill, B. The Impact of Algorithms on Judicial Discretion: Evidence from Regression Discontinuities.

[108] Crenshaw, K. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *The University of Chicago Legal Forum* (1989), 139.

[109] Crenshaw, K. W. Race, Reform, and Retrenchment: Transformation and Legitimation in Antidiscrimination Law. *Harvard Law Review 101*, 7 (1988), 1331–1387.

[110] Crespo, A. M. Systemic Facts: Toward Institutional Awareness in Criminal Courts. *Harvard Law Review 129* (2015), 2049–2117.

[111] Cullen, F. T., Jonson, C. L., and Nagin, D. S. Prisons Do Not Reduce Recidivism: The High Cost of Ignoring Science. *The Prison Journal 91*, 3_suppl (2011), 48S–65S.

[112] Cummings, M. L. Automation and Accountability in Decision Support System Interface Design. *Journal of Technology Studies* (2006).

[113] Cushing, T. 'Predictive Policing' Company Uses Bad Stats, Contractually-Obligated Shills To Tout Unproven 'Successes'. *Techdirt* (2013).

[114] Daley, D. J., and Vere-Jones, D. *An Introduction to the Theory of Point Processes*, vol. 1. Springer-Verlag New York, 2007.

[115] Danner, M. J., VanNostrand, M., and Spruance, L. M. Risk-Based Pretrial Release Recommendation and Supervision Guidelines. *Luminosity, Inc.* (2015).

[116] Daston, L., and Galison, P. *Objectivity*. Zone Books, 2007.

[117] Davis, A. Y. *Are Prisons Obsolete?* Seven Stories Press, 2003.

[118] Davis, L. M., Bozick, R., Steele, J. L., Saunders, J., and Miles, J. N. *Evaluating the Effectiveness of Correctional Education: A Meta-Analysis of Programs That Provide Education to Incarcerated Adults*. Rand Corporation, 2013.

[119] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting Depression via Social Media. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (2013), AAAI.

[120] DeFina, R., and Hannon, L. For incapacitation, there is no time like the present: The lagged effects of prisoner reentry on property and violent crime rates. *Social Science Research 39*, 6 (2010), 1004–1014.

[121] Delgado, R., and Stefancic, J. Images of the Outsider in American Law and Culture: Can Free Expression Remedy Systemic Social Ills. *Cornell Law Review 77*, 6 (1992), 1258–1297.

[122] Delgado, R., and Stefancic, J. *Critical Race Theory: An Introduction*, third ed. New York University Press, 2017.

[123] DeMichele, M., Baumgartner, P., Wenger, M., Barrick, K., Comfort, M., and Misra, S. The Public Safety Assessment: A Re-Validation and Assessment of Predictive Utility and Differential Prediction by Race and Gender in Kentucky.

[124] Desilver, D. Despite recent shootings, Chicago nowhere near US 'murder capital'. *Pew Research Center* (2014).

[125] Desmarais, S. L., Johnson, K. L., and Singh, J. P. Performance of Recidivism Risk Assessment Instruments in U.S. Correctional Settings. *Psychological Services 13*, 3 (2016), 206–222.

[126] Desmarais, S. L., and Singh, J. P. Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States.

[127] Development, U. R. Single Family Housing Repair Loans & Grants.

[128] Dewey, J. *Logic: The Theory of Inquiry*. H. Holt and Company, 1938.

[129] Dickinson, J., Díaz, M., Le Dantec, C. A., and Erete, S. "The Cavalry Ain't Coming in to Save Us": Supporting Capacities and Relationships Through Civic Tech. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 123:1–123:21.

[130] Dieterich, W., Mendoza, C., and Brennan, T. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe Inc. Research Department* (2016).

[131] Dietvorst, B. J., Simmons, J. P., and Massey, C. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General 144*, 1 (2015), 114–126.

[132] Dietvorst, B. J., Simmons, J. P., and Massey, C. Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science* (2016).

[133] Dobbie, W., Goldin, J., and Yang, C. S. The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review 108*, 2 (2018), 201–40.

[134] Doctorow, C. Algorithmic risk-assessment: hiding racism behind "empirical" black boxes. *Boing Boing* (2016).

[135] Doleac, J. L., and Hansen, B. The Unintended Consequences of "Ban the Box": Statistical Discrimination and Employment Outcomes When Criminal Histories Are Hidden. *Journal of Labor Economics 38*, 2 (2020), 321–374.

[136] Doshi-Velez, F., and Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[137] Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., and Wood, A. Accountability of AI Under the Law: The Role of Explanation. *arXiv preprint arXiv:1711.01134* (2017).

[138] Dourish, P. What we talk about when we talk about context. *Personal Ubiquitous Computing 8*, 1 (2004), 19–30.

[139] Dourish, P. Algorithms and their others: Algorithmic culture in context. *Big Data & Society 3*, 2 (2016).

[140] Downs, A. The law of peak-hour expressway congestion. *Traffic Quarterly 16*, 3 (1962), 393–409.

[141] Dressel, J., and Farid, H. The accuracy, fairness, and limits of predicting recidivism. *Science Advances 4*, 1 (2018).

[142] Dunne, A., and Raby, F. *Design Noir: The Secret Life of Electronic Objects*. Springer Science & Business Media, 2001.

[143] Duranton, G., and Turner, M. A. The Fundamental Law of Road Congestion: Evidence from US Cities. *The American Economic Review 101*, 6 (2011), 2616–2652.

[144] Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (2090255, 2012), ACM, pp. 214–226.

[145] Eberhardt, J. L., Davies, P. G., Purdie-Vaughns, V. J., and Johnson, S. L. Looking Deathworthy: Perceived Stereotypicality of Black Defendants Predicts Capital-Sentencing Outcomes. *Psychological Science 17*, 5 (2006), 383–386.

[146] Eckhouse, L., Lum, K., Conti-Cook, C., and Ciccolini, J. Layers of Bias: A Unified Approach for Understanding Problems With Risk Assessment. *Criminal Justice and Behavior 46*, 2 (2019), 185–209.

[147] EdBuild. $23 Billion.

[148] Edwards, L., and Veale, M. Slave to the Algorithm? Why a 'Right to an Explanation' Is Probably Not the Remedy You Are Looking For. *Duke Law & Technology Review. 16* (2017), 18–84.

[149] Elish, M. C. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society 5*, 0 (2019), 40–60.

[150] Elmalech, A., Sarne, D., Rosenfeld, A., and Erez, E. S. When Suboptimal Rules. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence* (2015), pp. 1313–1319.

[151] Englich, B., Mussweiler, T., and Strack, F. Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts' Judicial Decision Making. *Personality and Social Psychology Bulletin 32*, 2 (2006), 188–200.

[152] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., and Sandvig, C. "I Always Assumed That I Wasn't Really That Close to [Her]": Reasoning About Invisible Algorithms in News Feeds. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2015), CHI '15, ACM, pp. 153–162.

[153] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature 542* (2017), 115.

[154] Eubanks, V. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press, 2018.

[155] Fails, J. A., and Olsen, Jr., D. R. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (New York, NY, USA, 2003), IUI '03, ACM, pp. 39–45.

[156] Fang, L. In Her First Race, Kamala Harris Campaigned as Tough on Crime – And Unseated the Country's Most Progressive Prosecutor. *The Intercept* (2019).

[157] Farajtabar, M., Du, N., Gomez-Rodriguez, M., Valera, I., Zha, H., and Song, L. Shaping social activity by incentivizing users. In *Proceedings of The 28th Annual Conference on Neural Information Processing Systems* (2014), Curran Associates, Inc., pp. 2474–2482.

[158] Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and Removing Disparate Impact. In *21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), ACM, pp. 259–268.

[159] Felson, R. B. Routine activities and involvement in violence as actor, witness, or target. *Violence and Victims 12*, 3 (1997), 209–221.

[160] Ferguson, A. G. The Allure of Big Data Policing. *PrawfsBlawg* (2017).

[161] Ferguson, A. G. *The Rise of Big Data Policing: Surveillance, Race, and the Future of Law Enforcement*. NYU Press, 2017.

[162] Fiesler, C., Garrett, N., and Beard, N. What Do We Teach When We Teach Tech Ethics? A Syllabi Analysis. In *The 51st ACM Technical Symposium on Computer Science Education (SIGCSE '20)* (2020).

[163] Fisher, R. J. Social Desirability Bias and the Validity of Indirect Questioning. *Journal of Consumer Research 20*, 2 (1993), 303–315.

[164] Fisher, R. J., and Tellis, G. J. Removing Social Desirability Bias With Indirect Questioning: Is the Cure Worse Than the Disease? *Advances in Consumer Research 25* (1998), 563–567.

[165] Fishkin, J. *Bottlenecks: A New Theory of Equal Opportunity*. Oxford University Press, 2014.

[166] Flores, A. W., Bechtel, K., and Lowenkamp, C. T. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased against Blacks". *Federal Probation 80*, 2 (2016), 38–46.

[167] for Disease Control, T. C., and Prevention. Web-based Injury Statistics Query and Reporting System (WISQARS) Nonfatal Injury Reports, 2001 - 2013.

[168] for Progress, D. Polling The Left Agenda.

[169] Ford, M. A New Approach to Criminal-Justice Reform. *The Atlantic* (2015).

[170] Foster, D. NEW R package that makes XGBoost interpretable. *Medium: Applied Data Science* (2017).

[171] Frey, W. R., Patton, D. U., Gaskell, M. B., and McGregor, K. A. Artificial Intelligence and Inclusion: Formerly Gang-Involved Youth as Domain Experts for Analyzing Unstructured Twitter Data. *Social Science Computer Review* (2018), 0894439318788314.

[172] Friedman, B., and Nissenbaum, H. Bias in Computer Systems. *ACM Transactions on Information Systems 14*, 3 (1996), 330–347.

[173] Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics 29*, 2 (2001), 1189–1232.

[174] Fund, T. L. C. E. The Use of Pretrial "Risk Assessment" Instruments: A Shared Statement of Civil Rights Concerns.

[175] FWD.us. Broad, Bipartisan Support for Bold Pre-Trial Reforms in New York State.

[176] FWD.us. Every Second: The Impact of the Incarceration Crisis on America's Families.

[177] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumeé III, H., and Crawford, K. Datasheets for Datasets. *arXiv preprint arXiv:1803.09010* (2018).

[178] Ghandnoosh, N. Race and Punishment: Racial Perceptions of Crime and Support for Punitive Policies. *The Sentencing Project* (2014).

[179] Glaser, A., and Oremus, W. "A Collective Aghastness": Why Silicon Valley workers are demanding their employers stop doing business with the Trump administration. *Slate* (2018).

[180] Gneiting, T., and Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102*, 477 (2007), 359–378.

[181] Goel, S., Rao, J. M., and Shroff, R. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics 10*, 1 (2016), 365–394.

[182] Goff, P. A., Jackson, M. C., Leone, D., Lewis, B. A., Culotta, C. M., and DiTomasso, N. A. The Essence of Innocence: Consequences of Dehumanizing Black Children. *Journal of Personality and Social Psychology 106*, 4 (2014), 526–545.

[183] Goldmacher, S. Michael Bloomberg Pushed 'Stop-and-Frisk' Policing. Now He's Apologizing. *The New York Times* (2019).

[184] Gomez-Rodriguez, M., Leskovec, J., and Krause, A. Inferring networks of diffusion and influence. *ACM Transactions on Knowledge Discovery from Data 5*, 4 (2012).

[185] Gong, A. Ethics for powerful algorithms (1 of 4). *Medium* (2016).

[186] Goodwin, P., and Fildes, R. Judgmental Forecasts of Time Series Affected by Special Events: Does Providing a Statistical Forecast Improve Accuracy? *Journal of Behavioral Decision Making 12*, 1 (1999), 37–53.

[187] Gordon, E., and Walter, S. Meaningful Inefficiencies: Resisting the Logic of Technological Efficiency in the Design of Civic Systems. In *Civic Media: Technology, Design, Practice*, E. Gordon and P. Mihailidis, Eds. 2016, p. 243.

[188] Gorner, J. With violence up, Chicago police focus on a list of likeliest to kill, be killed. *Chicago Tribune* (2016).

[189] Gorz, A. *Strategy for Labor*. Beacon Press, 1967.

[190] Gottfredson, D. M. Effects of Judges' Sentencing Decisions on Criminal Careers.

[191] Green, B. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *arXiv preprint arXiv:1811.03435* (2018).

[192] Green, B. 'Fair' Risk Assessments: A Precarious Approach for Criminal Justice Reform. In *5th Workshop on Fairness, Accountability, and Transparency in Machine Learning* (2018).

[193] Green, B. Putting the J(ustice) in FAT. *Berkman Klein Center Collection - Medium* (2018).

[194] Green, B. *The Smart Enough City: Putting Technology in Its Place to Reclaim Our Urban Future*. MIT Press, 2019.

[195] Green, B. The False Promise of Risk Assessments: Epistemic Reform and the Limits of Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2020), FAT* '20, ACM.

[196] Green, B., and Chen, Y. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), FAT* '19, ACM, pp. 90–99.

[197] Green, B., and Chen, Y. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 50:1–50:24.

[198] Green, B., and Chen, Y. Algorithmic risk assessments can distort human decision-making in high-stakes government contexts.

[199] Green, B., Horel, T., and Papachristos, A. V. Modeling Contagion Through Social Networks to Explain and Predict Gunshot Violence in Chicago, 2006 to 2014. *JAMA Internal Medicine 177*, 3 (2017), 326–333.

[200] Green, B., and Hu, L. The Myth in the Methodology: Towards a Recontextualization of Fairness in Machine Learning. In *Machine Learning: The Debates workshop at the 35th International Conference on Machine Learning* (2018).

[201] Green, B., and Viljoen, S. Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2020), FAT* '20, Association for Computing Machinery, p. 19–31.

[202] Greene, D., Hoffmann, A. L., and Stark, L. Better, Nicer, Clearer, Fairer: A Critical Assessment of the Movement for Ethical Artificial Intelligence and Machine Learning. In *Proceedings of the 52nd Hawaii International Conference on System Sciences* (2019), pp. 2122–2131.

[203] Grenfell, B., Bjørnstad, O., and Kappey, J. Travelling waves and spatial hierarchies in measles epidemics. *Nature 414*, 6865 (2001), 716–723.

[204] Grewal, D. S., Kapczynski, A., and Purdy, J. Law and Political Economy: Toward a Manifesto. *Law and Political Economy* (2017).

[205] Griffin, D. 'Do Not Resist' traces the militarization of police with unprecedented access to raids and unrest. *Baltimore City Paper* (2016).

[206] Griffiths, E., and Chavez, J. M. Communities, street guns, and homicide trajectories in Chicago, 1980-1995: Merging methods for examing homicide trends across space and time. *Criminology 42*, 4 (2004), 941–978.

[207] Guthrie, C., Rachlinski, J. J., and Wistrich, A. J. Inside the Judicial Mind. *Cornell Law Review 86* (2000), 777.

[208] Hale, R. L. Coercion and Distribution in a Supposedly Non-Coercive State. *Political Science Quarterly 38*, 3 (1923), 470–494.

[209] Hall, N. Two concepts of causation. *Causation and Counterfactuals* (2004), 225–276.

[210] Hannah-Moffat, K., Maurutto, P., and Turnbull, S. Negotiated Risk: Actuarial Illusions and Discretion in Probation. *Canadian Journal of Law & Society/La Revue Canadienne Droit et Société 24*, 3 (2009), 391–409.

[211] Hannon, L., and DeFina, R. Violent Crime in African American and White Neighborhoods: Is Poverty's Detrimental Effect Race-Specific? *Journal of Poverty 9*, 3 (2005), 49–67.

[212] Haraway, D. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies 14*, 3 (1988), 575–599.

[213] Harcourt, B. E. Risk as a Proxy for Race: The Dangers of Risk Assessment. *Federal Sentencing Reporter 27*, 4 (2015), 237–243.

[214] Harding, S. *Is Science Multicultural?: Postcolonialisms, Feminisms, and Epistemologies*. Indiana University Press, 1998.

[215] Hardt, M., Price, E., and Srebro, N. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (2016), pp. 3315–3323.

[216] Harrington, C., Erete, S., and Piper, A. M. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proceedings of the ACM on Human-Computer Interaction 3*, CSCW (2019), 216:1–216:25.

[217] Harris, K., and Paul, R. Pretrial Integrity and Safety Act of 2017. *115th Congress* (2017).

[218] Hassein, N. Against Black Inclusion in Facial Recognition. *Digital Talking Drum* (2017).

[219] Hatzenbuehler, M. L., Keyes, K., Hamilton, A., Uddin, M., and Galea, S. The Collateral Damage of Mass Incarceration: Risk of Psychiatric Morbidity Among Nonincarcerated Residents of High-Incarceration Neighborhoods. *American Journal of Public Health 105*, 1 (2015), 138–143.

[220] Hawkes, A. G. Spectra of some self-exciting and mutually exciting point processes. *Biometrika 58*, 1 (1971), 83–90.

[221] Hawkins, A. J. Deadly Boeing Crashes Raise Questions About Airplane Automation. *The Verge* (2019).

[222] Haynie, D. L. Delinquent peers revisited: Does network structure matter? *American Journal of Sociology 106*, 4 (2001), 1013–1057.

[223] Haynie, D. L. Friendship networks and delinquency: The relative nature of peer delinquency. *Journal of Quantitative Criminology 18*, 2 (2002), 99–134.

[224] Heaton, P., Mayson, S., and Stevenson, M. The downstream consequences of misdemeanor pretrial detention. *Stanford Law Review 69* (2017), 711–794.

[225] Heller, S. B. Summer jobs reduce violence among disadvantaged youth. *Science 346*, 6214 (2014), 1219–1223.

[226] Hellman, D. Measuring Algorithmic Fairness. *Virginia Law Review* (2019).

[227] Hemenway, D. *Private Guns, Public Health*, vol. 498. University of Michigan Press, 2004.

[228] Hinton, E. K., Henderson, L., and Reed, C. An Unjust Burden: The Disparate Treatment of Black Americans in the Criminal Justice System.

[229] Hoffmann, A. L. Data Violence and How Bad Engineering Choices Can Damage Society. *Medium* (2018).

[230] Hoffmann, A. L. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information, Communication & Society 22*, 7 (2019), 900–915.

[231] Hohfeld, W. N. Some Fundamental Legal Conceptions as Applied in Judicial Reasoning. *The Yale Law Journal 23*, 1 (1913), 16–59.

[232] Holder, E. Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting and 13th State Criminal Justice Network Conference.

[233] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. *arXiv preprint arXiv:1805.03677* (2018).

[234] Holmes, O. W. The Path of the Law. *Harvard Law Review 10* (1897), 457–478.

[235] Honig, E. Elie Honig to Judge Grant.

[236] Horan, C. D. *Actuarial age: insurance and the emergence of neoliberalism in the postwar United States*. PhD thesis, 2011.

[237] Horvitz, E. Principles of Mixed-initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 1999), CHI '99, ACM, pp. 159–166.

[238] Hunt, P., Saunders, J., and Hollywood, J. S. *Evaluation of the Shreveport Predictive Policing Experiment*. RAND Corporation, 2014.

[239] Hutchinson, B., and Mitchell, M. 50 Years of Test (Un)fairness: Lessons for Machine Learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), Association for Computing Machinery, p. 49–58.

[240] Hutson, M. Artificial intelligence could identify gang crimes—and ignite an ethical firestorm. *Science* (2018).

[241] Hvistendahl, M. Can 'predictive policing' prevent crime before it happens? *Science* (2016).

[242] IBM. Predictive Analytics – Police Use Analytics to Reduce Crime.

[243] Inc., N. COMPAS Risk & Need Assessment System.

[244] Ingolfsson, A., Budge, S., and Erkut, E. Optimal ambulance location with random delays and travel times. *Health Care Management Science 11* (2008), 262–274.

[245] Institute, P. J. Pretrial Risk Assessment Can Produce Race-Neutral Results.

[246] Jackman, T. U.S. police chiefs group apologizes for 'historical mistreatment' of minorities. *The Washington Post* (2016).

[247] Jacobson, J., Dobbs-Marsh, J., Liberman, V., and Minson, J. A. Predicting Civil Jury Verdicts: How Attorneys Use (and Misuse) a Second Opinion. *Journal of Empirical Legal Studies 8* (2011), 99–119.

[248] Jagtenberg, C., Bhulai, S., and van der Mei, R. Optimal Ambulance Dispatching. In *Markov Decision Processes in Practice*. Springer, 2017, pp. 269–291.

[249] Jasanoff, S. *Risk Management and Political Culture*. Russell Sage Foundation, 1986.

[250] Jasanoff, S. The idiom of co-production. In *States of Knowledge: The Co-Production of Science and the Social Order*, S. Jasanoff, Ed. Routledge, 2004, ch. 1, pp. 1–12.

[251] Jasanoff, S. Ordering knowledge, ordering society. In *States of Knowledge: The Co-Production of Science and the Social Order*, S. Jasanoff, Ed. Routledge, 2004, pp. 13–45.

[252] Jasanoff, S. Making Order: Law and Science in Action. In *The Handbook of Science and Technology Studies*, E. J. Hackett, O. Amsterdamska, M. E. Lynch, and J. Wajcman, Eds., third ed. MIT Press, 2007, pp. 761–786.

[253] Jasanoff, S. *Designs on Nature: Science and Democracy in Europe and the United States*. Princeton University Press, 2011.

[254] Jasanoff, S. Future Imperfect: Science, Technology, and the Imaginations of Modernity. In *Dreamscapes of Modernity: Sociotechnical Imaginaries and the Fabrication of Power*, S. Jasanoff and S.-H. Kim, Eds. University of Chicago Press, 2015, ch. 1, pp. 1–47.

[255] Jensen, T., and Tilley, J. HB 463 – Statement from the Sponsors. *Criminal Law Reform: The First Year of HB 463* (2012).

[256] Joh, E. The Undue Influence of Surveillance Technology Companies on Policing. *New York University Law Review* (2017).

[257] Jouvenal, J. Police are using software to predict crime. Is it a 'holy grail' or biased against minorities? *The Washington Post* (2016).

[258] Kahn, J. *Race on the Brain: What Implicit Bias Gets Wrong About the Struggle for Racial Justice*. Columbia University Press, 2017.

[259] Kahneman, D. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

[260] Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., and Shapiro, J. R. Presumed Fair: Ironic Effects of Organizational Diversity Structures. *Journal of Personality and Social Psychology 104*, 3 (2013), 504–519.

[261] Kamar, E. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (2016), pp. 4070–4073.

[262] Kamar, E., Hacker, S., and Horvitz, E. Combining Human and Machine Intelligence in Large-scale Crowdsourcing. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems - Volume 1* (Richland, SC, 2012), AAMAS '12, International Foundation for Autonomous Agents and Multiagent Systems, pp. 467–474.

[263] Karakatsanis, A. Policing, Mass Imprisonment, and the Failure of American Lawyers. *Harvard Law Review Forum 128* (2015), 253.

[264] Karakatsanis, A. The Punishment Bureaucracy: How to Think About "Criminal Justice Reform". *The Yale Law Journal Forum 128* (2019), 848–935.

[265] Kazansky, B., Torres, G., van der Velden, L., Wissenbach, K., and Milan, S. Data for the Social Good: Toward a Data-Activist Research Agenda. *Good Data 4* (2019), 244.

[266] Kehl, D., Guo, P., and Kessler, S. Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing. *Responsive Communities Initiative, Berkman Klein Center for Internet & Society* (2017).

[267] Kennedy, D. The Critique of Rights in Critical Legal Studies. In *Left Legalism/Left Critique*, W. Brown and J. Halley, Eds. Duke University Press, 2002, pp. 178–228.

[268] Kennedy, D. Three Globalizations of Law and Legal Thought: 1850-2000. In *The New Law and Economic Eevelopment: A Critical Appraisal*, D. M. Trubek and A. Santos, Eds. 2006, pp. 19–73.

[269] Kennedy, D., and III, W. W. F. *The Canon of American Legal Thought*. Princeton University Press, 2006.

[270] Kennedy, D. M., Braga, A. A., and Piehl, A. M. The (un)known universe: Mapping gangs and gang violence in Boston. In *In: D. Weisburd and T. McEwen (Eds.), Crime Mapping and Crime Prevention* (1997), Citeseer.

[271] Kim, S., Kalev, A., and Dobbin, F. Progressive Corporations at Work: The Case of Diversity Programs. *NYU Review of Law and Social Change 36* (2012), 171.

[272] King, R. D., and Johnson, B. D. A Punishing Look: Skin Tone and Afrocentric Features in the Halls of Justice. *American Journal of Sociology 122*, 1 (2016), 90–124.

[273] Kirk, D. S. Examining the divergence across self-report and official data sources on inferences about the adolescent life-course of crime. *Journal of Quantitative Criminology 22*, 2 (2006), 107–129.

[274] Kiviat, B. The Moral Limits of Predictive Practices: The Case of Credit-Based Insurance Scores. *American Sociological Review 84*, 6 (2019), 1134–1158.

[275] Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., and Mullainathan, S. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics 133*, 1 (2018), 237–293.

[276] Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. Prediction Policy Problems. *American Economic Review 105*, 5 (2015), 491–95.

[277] Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. Discrimination in the Age of Algorithms. *Journal of Legal Analysis 10* (2019), 113–174.

[278] Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[279] Kleinberg, J., and Tardos, E. *Algorithm Design*. Pearson Education, Inc., 2006.

[280] Klockars, C. B. Some Really Cheap Ways of Measuring What Really Matters.

[281] Koepke, J. L., and Robinson, D. G. Danger Ahead: Risk Assessment and the Future of Bail Reform. *Washington Law Review 93* (2018), 1725–1807.

[282] Koester, S., Glanz, J., and Barón, A. Drug sharing among heroin networks: Implications for HIV and hepatitis B and C prevention. *AIDS and Behavior 9*, 1 (2005), 27–39.

[283] Kortylewski, A., Egger, B., Schneider, A., Gerig, T., Morel-Forster, A., and Vetter, T. Analyzing and Reducing the Damage of Dataset Bias to Face Recognition With Synthetic Data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (2019).

[284] Krivo, L. J., Peterson, R. D., and Kuhl, D. C. Segregation, Racial Structure, and Neighborhood Violent Crime. *American Journal of Sociology 114*, 6 (2009), 1765–1802.

[285] Kulesza, T., Amershi, S., Caruana, R., Fisher, D., and Charles, D. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 3075–3084.

[286] Kushner, R. Is Prison Necessary? Ruth Wilson Gilmore Might Change Your Mind. *The New York Times* (2019).

[287] Lai, V., and Tan, C. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2019), FAT* '19, ACM, pp. 29–38.

[288] Langley, P. The changing science of machine learning. *Machine Learning 82*, 3 (2011), 275–279.

[289] Larson, J., Mattu, S., Kirchner, L., and Angwin, J. How We Analyzed the COMPAS Recidivism Algorithm. *ProPublica* (2016).

[290] Latour, B. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, 1988.

[291] Laura, and Foundation, J. A. Public Safety Assessment: Risk Factors and Formula.

[292] Laura, and Foundation, J. A. Guide to the Release Conditions Matrix.

[293] Laura, and Foundation, J. A. Public Safety Assessment (PSA) - Intro.

[294] Lavigne, S., Clifton, B., and Tseng, F. Predicting Financial Crime: Augmenting the Predictive Policing Arsenal. *The New Inquiry* (2017).

[295] Lee, J. D., and See, K. A. Trust in Automation: Designing for Appropriate Reliance. *Human Factors 46*, 1 (2004), 50–80.

[296] Lee, M. K. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society 5*, 1 (2018), 2053951718756684.

[297] Leith, P. *Formalism in AI and Computer Science*. Ellis Horwood, 1990.

[298] Leskovec, J., McGlohon, M., Faloutsos, C., Glance, N. S., and Hurst, M. Patterns of cascading behavior in large blog graphs. In *Proceedings of the 2007 SIAM International Conference on Data Mining* (2007), SIAM, pp. 551–556.

[299] Lessig, L. *Code*. Basic Books, 2009.

[300] Levashina, J., Hartwell, C. J., Morgeson, F. P., and Campion, M. A. The Structured Employment Interview: Narrative and Quantitative Review of the Research Literature. *Personnel Psychology 67*, 1 (2014), 241–293.

[301] Leventhal, G. S. What Should Be Done with Equity Theory? In *Social Exchange*. Springer, 1980, pp. 27–55.

[302] Lim, J. S., and O'Connor, M. Judgemental Adjustment of Initial Forecasts: Its Effectiveness and Biases. *Journal of Behavioral Decision Making 8*, 3 (1995), 149–168.

[303] Linderman, S., and Adams, R. Discovering Latent Network Structure in Point Process Data. In *Proceedings of the 31st International Conference on Machine Learning* (2014), JMLR.

[304] Lininger, T. *Multivariate Hawkes Processes*. Dissertation, 2009.

[305] LLC, E. Who's Behind ICE? The Tech and Data Companies Fueling Deportations. *Mijente* (2018).

[306] Llewellyn, K. N. A Realistic Jurisprudence–The Next Step. *Columbia Law Review 30* (1930), 431.

[307] Llewellyn, K. N. Some Realism about Realism: Responding to Dean Pound. *Harvard Law Review 44*, 8 (1931), 1222–1264.

[308] Lochner, L., and Moretti, E. The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports. *American Economic Review 94*, 1 (2004), 155–189.

[309] Logg, J. M., Minson, J. A., and Moore, D. A. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes 151* (2019), 90 – 103.

[310] Lopez, G. Amy Klobuchar's record as a "tough on crime" prosecutor, explained. *Vox* (2019).

[311] Lowenkamp, C. T., and Whetzel, J. The Development of an Actuarial Risk Assessment Instrument for U.S. Pretrial Services. *Federal Probation 73* (2009).

[312] Lum, K. Predictive Policing Reinforces Police Bias.

[313] Lum, K., and Isaac, W. To predict and serve? *Significance 13*, 5 (2016), 14–19.

[314] Lum, K., Ma, E., and Baiocchi, M. The causal impact of bail on case outcomes for indigent defendants in New York City. *Observational Studies 3* (2017), 39–64.

[315] Lynch, M. *Hard Bargains: The Coercive Power of Drug Laws in Federal Court*. Russell Sage Foundation, 2016.

[316] MacKinnon, C. A. Feminism, Marxism, Method, and the State: An Agenda for Theory. *Signs: Journal of Women in Culture and Society 7*, 3 (1982), 515–544.

[317] MacKinnon, C. A. Substantive Equality: A Perspective. *Minnesota Law Review 96* (2011).

[318] Mahony, M. The predictive state: Science, territory and the future of the Indian climate. *Social Studies of Science 44*, 1 (2014), 109–133.

[319] Main, F. Cook County judges not following bail recommendations: study. *Chicago Sun-Times* (2016).

[320] Makowski, D., Ben-Shachar, M. S., Chen, S. H. A., and Lüdecke, D. Indices of Effect Existence and Significance in the Bayesian Framework. *Frontiers in Psychology 10*, 2767 (2019).

[321] Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. *Journal of Open Source Software 4*, 40 (2019), 1541.

[322] Marsan, D., and Lengliné, O. Extending earthquakes' reach through cascading. *Science 319*, 5866 (2008), 1076–1079.

[323] Mateescu, A., and Elish, M. C. AI in Context: The Labor of Integrating New Technologies. *Data & Society* (2019).

[324] Matias, J. N., and Mou, M. CivilServant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (2018), CHI '18, ACM, pp. 9:1–9:13.

[325] Maurutto, P., and Hannah-Moffat, K. Understanding risk in the context of the Youth Criminal Justice Act. *Canadian Journal of Criminology and Criminal Justice 49*, 4 (2007), 465–491.

[326] Mayson, S. G. Dangerous Defendants. *Yale Law Journal 127*, 3 (2018), 490–568.

[327] Mayson, S. G. Bias In, Bias Out. *Yale Law Journal 128*, 8 (2019), 2218–2300.

[328] McDowall, D., and Curtis, K. M. Seasonal variation in homicide and assault across large US cities. *Homicide Studies* (2014).

[329] McDowall, D., Loftin, C., and Pate, M. Seasonal cycles in crime, and their variability. *Journal of Quantitative Criminology 28*, 3 (2012), 389–410.

[330] McGloin, J. M., and Piquero, A. R. On the relationship between co-offending network redundancy and offending versatility. *Journal of Research in Crime and Delinquency* (2009).

[331] McLeod, A. M. Confronting Criminal Law's Violence: The Possibilities of Unfinished Alternatives. *Unbound: Harvard Journal of the Legal Left 8* (2013), 109–132.

[332] McLeod, A. M. Prison Abolition and Grounded Justice. *UCLA Law Review 62* (2015), 1156–1239.

[333] McLeod, A. M. Envisioning Abolition Democracy. *Harvard Law Review 132* (2019), 1613–1649.

[334] Meehl, P. E. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, 1954.

[335] Meng, A., and DiSalvo, C. Grassroots resource mobilization through counter-data action. *Big Data & Society 5*, 2 (2018).

[336] Merler, M., Ratha, N., Feris, R. S., and Smith, J. R. Diversity in Faces. *arXiv preprint arXiv:1901.10436* (2019).

[337] Metcalf, J., Moss, E., and boyd, d. Owning Ethics: Corporate Logics, Silicon Valley, and the Institutionalization of Ethics. *Social Research 86*, 2 (2019), 449–476.

[338] Mijente. 1,200+ Students at 17 Universities Launch Campaign Targeting Palantir.

[339] Milgram, A. Why smart statistics are the key to fighting crime. *TED* (2014).

[340] Miller, A. P. Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review* (2018).

[341] Minow, M. *Making All the Difference: Inclusion, Exclusion, and American Law*. Cornell University Press, 1991.

[342] Minow, M. The Path as Prologue. *Harvard Law Review 110*, 5 (1997), 1023–1027.

[343] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), ACM, pp. 220–229.

[344] Moats, D., and Seaver, N. "You Social Scientists Love Mind Games": Experimenting in the "divide" between data science and critical algorithm studies. *Big Data & Society 6*, 1 (2019), 2053951719833404.

[345] Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. Self-exciting point process modeling of crime. *Journal of the American Statistical Association 106*, 493 (2011), 100–108.

[346] Monahan, J., and Skeem, J. L. Risk Assessment in Criminal Sentencing. *Annual Review of Clinical Psychology 12* (2016), 489–513.

[347] Morenoff, J. D., and Sampson, R. J. Violent crime and the spatial dynamics of neighborhood transition: Chicago, 1970–1990. *Social Forces 76*, 1 (1997), 31–64.

[348] Morenoff, J. D., Sampson, R. J., and Raudenbush, S. W. Neighborhood inequality, collective efficacy, and the spatial dynamics of urban violence. *Criminology 39*, 3 (2001), 517–559.

[349] Morozov, E. *To Save Everything, Click Here: The Folly of Technological Solutionism*. PublicAffairs, 2014.

[350] Mosier, K. L., Dunbar, M., McDonnell, L., Skitka, L. J., Burdick, M., and Rosenblatt, B. Automation Bias and Errors: Are Teams Better than Individuals? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting 42*, 3 (1998), 201–205.

[351] Moskos, P. *Cop in the Hood: My Year Policing Baltimore's Eastern District*. Princeton University Press, 2008.

[352] Moyn, S. Civil Liberties and Endless War. *Dissent* (2015).

[353] Mozur, P. One Month, 500,000 Face Scans: How China Is Using A.I. to Profile a Minority. *The New York Times* (2019).

[354] Murakawa, N. *The First Civil Right: How Liberals Built Prison America*. Oxford University Press, 2014.

[355] myFICO. Understanding FICO Scores.

[356] Møller, J., and Rasmussen, J. G. Perfect Simulation of Hawkes Processes. *Advances in Applied Probability* (2005), 629–646.

[357] Narayanan, M., Chen, E., He, J., Kim, B., Gershman, S., and Doshi-Velez, F. How do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint arXiv:1802.00682* (2018).

[358] Neal, L. V. I., McCray, A. D., Webb-Johnson, G., and Bridgest, S. T. The Effects of African American Movement Styles on Teachers' Perceptions and Reactions. *The Journal of Special Education 37*, 1 (2003), 49–57.

[359] Neff, G., Tanweer, A., Fiore-Gartland, B., and Osburn, L. Critique and Contribute: A Practice-Based Framework for Improving Critical Data Studies and Data Science. *Big Data 5*, 2 (2017), 85–97.

[360] Network, D. J. Design Justice Network Principles.

[361] New Jersey Courts. One Year Criminal Justice Reform Report to the Governor and the Legislature.

[362] Nisbett, R. E., and Wilson, T. D. Telling More Than We Can Know: Verbal Reports on Mental Processes. *Psychological Review 84*, 3 (1977), 231–259.

[363] Nissenbaum, H. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press, 2009.

[364] Noble, S. U. Google search: Hyper-visibility as a means of rendering black women and girls invisible. *InVisible Culture*, 19 (2013).

[365] Noble, S. U. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press, 2018.

[366] Northpointe, Inc. Practitioner's Guide to COMPAS Core.

[367] Northpointe, Inc. Sample-COMPAS-Risk-Assessment-COMPAS-"CORE".

[368] Norton, P. D. *Fighting Traffic: The Dawn of the Motor Age in the American City*. MIT Press, 2011.

[369] Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science 366*, 6464 (2019), 447–453.

[370] Ochigame, R. The Long History of Algorithmic Fairness. *Phenomenal World* (2020).

[371] of Investigation, T. F. B. Crime in the United States.

[372] O'Neil, C. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. Broadway Books, 2017.

[373] Onuoha, M. Notes on Algorithmic Violence.

[374] Osgood, D. W., Wilson, J. K., O'Malley, P. M., Bachman, J. G., and Johnston, L. D. Routine activities and individual deviant behavior. *American Sociological Review 61*, 4 (1996), 635–655.

[375] Otis, G. A., and Dillon, N. Google using dubious tactics to target people with 'darker skin' in facial recognition project: sources. *New York Daily News* (2019).

[376] O'Malley, N. To predict and to serve: the future of law enforcement. *The Sydney Morning Herald* (2013).

[377] O'Neil, C. ProPublica report: recidivism risk models are racist. *mathbabe* (2016).

[378] Packer, G. Change the World. *The New Yorker* (2013).

[379] Pager, D. The Mark of a Criminal Record. *American Journal of Sociology 108*, 5 (2003), 937–975.

[380] Papachristos, A. V. 48 years of crime in Chicago: A descriptive analysis of serious crime trends from 1965 to 2013. *Yale University Institution for Social and Policy Studies Working Paper*, ISPS13-023 (2013).

[381] Papachristos, A. V., Braga, A. A., and Hureau, D. Social networks and the risk of gunshot injury. *Journal of Urban Health 89*, 6 (2012), 992–1003.

[382] Papachristos, A. V., Braga, A. A., Piza, E., and Grossman, L. The company you keep? The spillover effects of gang membership on individual gunshot victimization in a co-offending network. *Criminology 53*, 4 (2015), 624–649.

[383] Papachristos, A. V., and Kirk, D. S. Changing the street dynamic: Evaluating Chicago's Group Violence Reduction Strategy. *Criminology & Public Policy 14*, 3 (2015), 525–558.

[384] Papachristos, A. V., Meares, T. L., and Fagan, J. Attention felons: Evaluating Project Safe Neighborhoods in Chicago. *Journal of Empirical Legal Studies 4*, 2 (2007).

[385] Papachristos, A. V., Wildeman, C., and Roberto, E. Tragic, but not random: The social contagion of nonfatal gunshot injuries. *Social Science & Medicine 125*, 1 (2015).

[386] Passi, S., and Barocas, S. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), ACM, pp. 39–48.

[387] Passi, S., and Jackson, S. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (2017), ACM, pp. 2436–2447.

[388] Pearl, J. *Causality*. Cambridge University Press, 2009.

[389] Peller, G. Race Consciousness. *Duke Law Journal* (1990), 758.

[390] Peterson, R. D., and Krivo, L. J. *Divergent Social Worlds: Neighborhood Crime and the Racial-Spatial Divide*. Russell Sage, New York, 2010.

[391] Phillips, J. A. White, Black, and Latino Homicide Rates: Why the Difference? *Social Problems 49*, 3 (2014), 349–373.

[392] Pinch, T. J., and Bijker, W. E. The Social Construction of Facts and Artifacts: Or How the Sociology of Science and the Sociology of Technology Might Benefit Each Other. In *The Social Construction of Technological Systems*, W. E. Bijker, T. P. Hughes, and T. Pinch, Eds. MIT Press, 1987.

[393] Porrino, C. S. Attorney General Law Enforcement Directive 2016-6 v3.0.

[394] Porrino, C. S. Attorney General Law Enforcement Directive No. 2016-6 v2.0.

[395] Porter, T. M. *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton University Press, 1995.

[396] Posner, R. A. The Path Away from the Law. *Harvard Law Review 110* (1997), 1039.

[397] Potash, E., Brew, J., Loewi, A., Majumdar, S., Reece, A., Walsh, J., Rozier, E., Jorgenson, E., Mansour, R., and Ghani, R. Predictive Modeling for Public Health: Preventing Childhood Lead Poisoning. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2015), Association for Computing Machinery, p. 2039–2047.

[398] Pound, R. Liberty of Contract. *Yale Law Journal 18*, 7 (1909).

[399] Pound, R. Law in Books and Law in Action. *American Law Review 44*, 1 (1910), 12–36.

[400] Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. Manipulating and Measuring Model Interpretability. *arXiv preprint arXiv:1802.07810* (2018).

[401] PredPol. Proven Results of our Predictive Policing Software.

[402] PredPol. PredPol on Current TV with Santa Cruz Crime Analyst Zach Friend.

[403] Pretrial Justice Institute. The State of Pretrial Justice in America.

[404] Pretrial Justice Institute. Scan of Pretrial Practices. *Pretrial Justice Institute* (2019).

[405] Prins, S. J., and Reich, A. Can we avoid reductionism in risk reduction? *Theoretical Criminology* (2017), 1–21.

[406] Promise, G., Aid, T. N. L., Association, D., for Public Defense, T. N. A., and of Criminal Defense Lawyers, T. N. A. Joint Statement in Support of the Use of Pretrial Risk Assessment Instruments.

[407] Purtle, J., Dicker, R., Cooper, C., Corbin, T., Greene, M. B., Marks, A., Creaser, D., Topp, D., and Moreland, D. Hospital-based violence intervention programs save lives and money. *Journal of Trauma and Acute Care Surgery 75*, 2 (2013), 331–333.

[408] Quillian, L., and Pager, D. Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime. *American Journal of Sociology 107*, 3 (2001), 717–767.

[409] Rachlinski, J. J., Johnson, S. L., Wistrich, A. J., and Guthrie, C. Does Unconscious Racial Bias Affect Trial Judges. *Notre Dame Law Review 84* (2008), 1195–1246.

[410] Rachlinski, J. J., and Wistrich, A. J. Gains, Losses, and Judges: Framing and the Judiciary. *Notre Dame Law Review 94*, 2 (2018), 521–582.

[411] Raji, I. D., and Buolamwini, J. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (2019), ACM, pp. 429–435.

[412] Rasmussen, J. G. Temporal point processes: The conditional intensity function. Tech. rep., January 24, 2011 2011.

[413] Rausand, M. *Risk Assessment: Theory, Methods, and Applications*. John Wiley & Sons, 2013.

[414] Reiman, J., and Leighton, P. *The Rich Get Richer and the Poor Get Prison: Ideology, Class, and Criminal Justice*. Routledge, 2015.

[415] Reisman, D., Schultz, J., Crawford, K., and Whittaker, M. Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability.

[416] Review, H. L. Introduction. *Harvard Law Review 132*, 6 (2019), 1568–1574.

[417] Ribeiro, M. T., Singh, S., and Guestrin, C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 1135–1144.

[418] Richardson, R., Schultz, J. M., and Southerland, V. M. Litigating Algorithms 2019 US Report.

[419] Robinson, D., and Koepke, L. Stuck in a Pattern. *Upturn* (2016).

[420] Rose, D. R., and Clear, T. R. Incarceration, Social Capital, and Crime: Implications for Social Disorganization Theory. *Criminology 36*, 3 (1998), 441–480.

[421] Rose, K. The Making of a YouTube Radical. *The New York Times* (2019).

[422] Rosenberg, M., and Levinson, R. Trump's catch-and-detain policy snares many who have long called U.S. home. *Reuters* (2018).

[423] Rosenquist, J., Fowler, J. H., and Christakis, N. A. Social network determinants of depression. *Molecular Psychiatry 16*, 3 (2011), 273–281.

[424] Rothstein, R. *The Color of Law: A Forgotten History of How Our Government Segregated America*. Liveright Publishing Corporation, 2017.

[425] Rubin, J. Stopping crime before it starts. *Los Angeles Times* (2010).

[426] Sadowski, J., and Bendor, R. Selling Smartness: Corporate Narratives and the Smart City as a Sociotechnical Imaginary. *Science, Technology, & Human Values 44*, 3 (2019), 540–563.

[427] Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., and Ghani, R. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[428] Salman, J., Coz, E. L., and Johnson, E. Florida's broken sentencing system. *Sarasota Herald-Tribune* (2016).

[429] Sampson, R. J. *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press, 2012.

[430] Sampson, R. J. *Great American City: Chicago and the Enduring Neighborhood Effect*. University of Chicago Press, 2012.

[431] Sampson, R. J., Morenoff, J. D., and Raudenbush, S. Social Anatomy of Racial and Ethnic Disparities in Violence. *American Journal of Public Health 95*, 2 (2005), 224–232.

[432] Sampson, R. J., and Wilson, W. J. Toward a Theory of Race, Crime, and Urban Inequality. In *Crime and Inequality*, J. Hagan and R. Peterson, Eds., vol. 1995. 1995, pp. 37–54.

[433] Sarkar, P., Kortela, J., Boriouchkine, A., Zattoni, E., and Jämsä-Jounela, S.-L. Data-Reconciliation Based Fault-Tolerant Model Predictive Control for a Biomass Boiler. *Energies 10*, 2 (2017), 194.

[434] Saunders, J., Hunt, P., and Hollywood, J. S. Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology 12*, 3 (2016), 347–371.

[435] Schauer, F. Formalism. *Yale Law Journal 97*, 4 (1987), 509–548.

[436] Scheiber, N., and Conger, K. Uber and Lyft Drivers Gain Labor Clout, With Help From an App. *The New York Times* (2019).

[437] Schneier, B. *Click Here to Kill Everybody: Security and Survival in a Hyper-connected World.* WW Norton & Company, 2018.

[438] Schnell, C., Braga, A. A., and Piza, E. L. The Influence of Community Areas, Neighborhood Clusters, and Street Segments on the Spatial Variability of Violent Crime in Chicago. *Journal of Quantitative Criminology* (2016), 1–28.

[439] Schofield, W. Christopher Columbus Langdell. *The American Law Register 55*, 5 (1907), 273–296.

[440] Schuppe, J. Post Bail. *NBC News* (2017).

[441] Scott, J. C. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed.* Yale University Press, 1998.

[442] Scurich, N., and John, R. S. The Effect of Framing Actuarial Risk Probabilities on Involuntary Civil Commitment Decisions. *Law and Human Behavior 35*, 2 (2011), 83–91.

[443] Seaver, N. The nice thing about context is that everyone has it. *Media, Culture & Society 37*, 7 (2015), 1101–1109.

[444] Seaver, N. Algorithms as culture: Some tactics for the ethnography of algorithmic systems. *Big Data & Society 4*, 2 (2017), 2053951717738104.

[445] Seaver, N. Knowing Algorithms. In *digitalSTS: A Field Guide for Science & Technology Studies*, J. Vertesi and D. Ribes, Eds. 2019, pp. 412–422.

[446] Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (2019), ACM, pp. 59–68.

[447] Sen, A., and Smith, T. *Gravity Models of Spatial Interaction Behavior.* Springer Science & Business Media, 2012.

[448] Shalizi, C. R., and Thomas, A. C. Homophily and contagion are generically confounded in observational social network studies. *Sociological methods & research 40*, 2 (2011), 211–239.

[449] Shalom, A., Tvedt, C., Krakora, J. E., and Price, D. D. The New Jersey Pretrial Justice Manual.

[450] Siddiqi, N. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* Wiley, 2012.

[451] Skeem, J., Scurich, N., and Monahan, J. Impact of Risk Assessment on Judges' Fairness in Sentencing Relatively Poor Defendants. *Law & Human Behavior* (2019).

[452] Skeem, J. L., and Lowenkamp, C. T. Risk, Race, & Recidivism: Predictive Bias and Disparate Impact. *Criminology 54*, 4 (2016), 680–712.

[453] Slutkin, G. *Violence is a contagious disease.* National Acadamy of Sciences, 2013, pp. 94–111.

[454] Smith, B. C. The limits of correctness. *ACM SIGCAS Computers and Society 14* (1985), 18–26.

[455] Smith, J. 'Minority Report' Is Real — And It's Really Reporting Minorities. *Mic* (2015).

[456] Smith, J. U.S. Courts Are Using Algorithms Riddled With Racism to Hand Out Sentences. *Mic* (2016).

[457] Smyth, T., and Dimond, J. Anti-oppressive design. *Interactions 21*, 6 (2014), 68–71.

[458] Sommers, R. Will Putting Cameras on Police Reduce Polarization? *Yale Law Journal 125*, 5 (2016), 1304–1362.

[459] Springer, A., Hollis, V., and Whittaker, S. Dice in the black box: User experiences with an inscrutable algorithm. In *2017 AAAI Spring Symposium Series* (2017).

[460] Starr, S. B. Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review 66*, 4 (2014), 803–872.

[461] Steinhart, D. Juvenile detention risk assessment: A practice guide to juvenile detention reform. *The Annie E. Casey Foundation* (2006).

[462] Stevenson, B. Why American Prisons Owe Their Cruelty to Slavery. *The New York Times Magazine* (2019).

[463] Stevenson, M. T. Risk Assessment: The Devil's in the Details. *The Crime Report* (2017).

[464] Stevenson, M. T. Assessing Risk Assessment in Action. *Minnesota Law Review 103* (2018), 303–384.

[465] Stevenson, M. T., and Doleac, J. L. The Roadblock to Reform. *The American Constitution Society* (2018).

[466] Stevenson, M. T., and Doleac, J. L. Algorithmic Risk Assessment in the Hands of Humans. *Available at SSRN* (2019).

[467] Stewart, J. Q. Demographic gravitation: evidence and applications. *Sociometry* (1948), 31–58.

[468] Strickland, E. How IBM Watson Overpromised and Underdelivered on AI Health Care. *IEEE Spectrum* (2019).

[469] Suchman, L., Blomberg, J., Orr, J. E., and Trigg, R. Reconstructing Technologies as Social Practice. *American Behavioral Scientist 43*, 3 (1999), 392–408.

[470] Sullivan, C. M., and O'Keeffe, Z. P. Evidence that curtailing proactive policing can reduce major crime. *Nature Human Behaviour 1* (2017), 730–737.

[471] Sunstein, C. R. Algorithms, Correcting Biases. *Social Research 86*, 2 (2019), 499–511.

[472] Sutherland, E. H. *Principles of Criminology (4th edition)*, fourth edition ed. J.B. Lippincott, Philadephia, PA, 1947.

[473] Sweeney, L. Discrimination in Online Ad Delivery. *Queue 11*, 3 (2013), 10–29.

[474] Sylvester, J., and Raff, E. What About Applied Fairness? In *Machine Learning: The Debates Workshop at the 35th International Conference on Machine Learning* (2018).

[475] Tachet, R., Santi, P., Sobolevsky, S., Reyes-Castro, L. I., Frazzoli, E., Helbing, D., and Ratti, C. Revisiting Street Intersections Using Slot-Based Systems. *PLOS ONE 11*, 3 (2016), e0149607.

[476] Tett, G. Mapping crime - or stirring hate? *Financial Times* (2014).

[477] Thornberry, T. P., and Krohn, M. D. Comparison of self-report and official data for measuring crime. In *Measurement problems in criminal justice research: Workshop summary* (2002), The National Academic Press Washington, DC, pp. 43–94.

[478] Tita, G. E., Cohen, J., and Engberg, J. An Ecological Study of the Location of Gang "Set Space". *Social Problems 52*, 2 (2005), 272–299.

[479] Toyama, K. *Geek Heresy: Rescuing Social Change from the Cult of Technology*. PublicAffairs, 2015.

[480] Tracy, M., Braga, A. A., and Papachristos, A. V. The Transmission of Gun and Other Weapon-Involved Violence Within Social Networks. *Epidemiologic Reviews* (2016).

[481] Turanovic, J. J., and Young, J. T. N. Violent offending and victimization in adolescence: Social network mechanisms and homophily. *Criminology* (2016), n/a–n/a.

[482] Tushnet, M. An Essay on Rights. *Texas Law Review 62*, 8 (1983), 1363–1403.

[483] Tushnet, M. The Critique of Rights. *SMU Law Review 47* (1993), 23–34.

[484] Tuttle, C. Snapping Back: Food Stamp Bans and Criminal Recidivism. *American Economic Journal: Economic Policy 11*, 2 (2019), 301–27.

[485] Tversky, A., and Kahneman, D. The Framing of Decisions and the Psychology of Choice. *Science 211*, 4481 (1981), 453–458.

[486] Unger, R. M. *False Necessity: Anti-Necessitarian Social Theory in the Service of Radical Democracy*. Cambridge University Press, 1987.

[487] United States Department of Justice Federal Bureau of Investigation. Murder Offenders by Age, Sex, Race, and Ethnicity, 2017. *Crime in the United States* (2018).

[488] United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. State Court Processing Statistics, 1990-2009: Felony Defendants in Large Urban Counties, 2014.

[489] United States Sentencing Commission. Sentencing Guidelines and Policy Statements.

[490] U.S. Supreme Court. *Lochner v. New York*. 198 U.S. 45, 1905.

[491] U.S. Supreme Court. *McCleskey v. Kemp*. 481 U.S. 279, 1987.

[492] U.S. Supreme Court. *United States v. Salerno*. 481 U.S. 739, 1987.

[493] Ustun, B., Spangher, A., and Liu, Y. Actionable Recourse in Linear Classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (New York, NY, USA, 2019), FAT* '19, ACM, pp. 10–19.

[494] Ventures, A. Public Safety Assessment FAQs ("PSA 101").

[495] Ventures, A. Statement of Principles on Pretrial Justice and Use of Pretrial Risk Assessment.

[496] Viboud, C., Bjørnstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., and Grenfell, B. T. Synchrony, waves, and spatial hierarchies in the spread of influenza. *Science 312*, 5772 (2006), 447–451.

[497] Vieraitis, L. M., Kovandzic, T. V., and Marvell, T. B. The Criminogenic Effects of Imprisonment: Evidence from State Panel Data, 1974–2002. *Criminology & Public Policy 6*, 3 (2007), 589–622.

[498] Vincent, G. M., Guy, L. S., and Grisso, T. Risk Assessment in Juvenile Justice: A Guidebook for Implementation.

[499] Vincent, J. Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge* (2016).

[500] Visvanathan, S. Knowledge, justice and democracy. In *Science and Citizens: Globalization and the Challenge of Engagement.*, M. Leach, I. Scoones, and B. Wynne, Eds. Zed Books, 2005.

[501] Vitale, A. S. *The End of Policing*. Verso Books, 2017.

[502] Warr, M. *Companions in Crime: The Social Aspects of Criminal Conduct*. Cambridge University Press, New York, 2002.

[503] Watch, H. R. "Not in it for Justice": How California's Pretrial Detention and Bail System Unfairly Punishes Poor People.

[504] Watts, D. J., and Strogatz, S. H. Collective dynamics of 'small-world' networks. *Nature 393*, 6684 (1998), 440–442.

[505] Weir, A. Formalism in the Philosophy of Mathematics. In *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2015.

[506] Western, B. *Punishment and Inequality in America*. Russell Sage Foundation, 2006.

[507] Wexler, R. Code of Silence. *Washington Monthly* (2017).

[508] Wexler, R. Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System. *Stanford Law Review 70* (2018), 1343–1429.

[509] Wildeman, C., and Wang, E. A. Mass incarceration, public health, and widening inequality in the USA. *The Lancet 389*, 10077 (2017), 1464–1474.

[510] Wilson, T., and Murgia, M. Uganda confirms use of Huawei facial recognition cameras. *Financial Times* (2019).

[511] Wing, J. Computational Thinking—What and Why? *The Link Magazine* (2011), 20–23.

[512] Winner, L. *The Whale and the Reactor: A Search for Limits in an Age of High Technology*. University of Chicago Press, 1986.

[513] Wintemute, G. J. The epidemology of firearm violence in the twenty-first century United States. *Annual Review of Public Health 36* (2014), 8.1–8.15.

[514] Wisconsin Supreme Court. *State v. Loomis*. 881 Wis. N.W.2d 749, 2016.

[515] Wood, G., and Papachristos, A. V. Reducing gunshot victimization in high-risk social networks through direct and spillover effects. *Nature Human Behaviour 3*, 11 (2019), 1164–1170.

[516] Wu, J. Optimize What? *Commune* (2019).

[517] Yan, Q., Kao, M., and Barrera, M. Algorithm-in-the-Loop with Plant Model Simulation, Reusable Test Suite in Production Codes Verification and Controller Hardware-in-the-Loop Bench Testing. *SAE Technical Paper*, 0148-7191 (2010).

[518] Yang, C. S. Toward an Optimal Bail System. *New York University Law Review 92*, 5 (2017), 1399–1493.

[519] Yaniv, I. Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes 93*, 1 (2004), 1–13.

[520] Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. Making sense of recommendations. *Journal of Behavioral Decision Making* (2019).

[521] Young, J. T. How do they 'end up together'? A social network analysis of self-control, homophily, and adolescent relationships. *Journal of Quantitative Criminology 27*, 3 (2011), 251–273.

[522] Zacka, B. *When the State Meets the Street: Public Service and Moral Agency*. Harvard University Press, 2017.

[523] Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. Defending Against Neural Fake News. *arXiv preprint arXiv:1905.12616* (2019).

[524] Zeoli, A. M., Pizzaro, J. M., Grady, S. C., and Melde, C. Homicide as an infectious disease: Using public health methods to investigate the diffusion of homicide. *Justice Quarterly* (2012).

[525] Zhen, L., Wang, K., Hu, H., and Chang, D. A simulation optimization framework for ambulance deployment and relocation problems. *Computers & Industrial Engineering 72* (2014), 12–23.

[526] Zimmerman, E. Teachers Are Turning to AI Solutions for Assistance. *EdTech Magazine* (2018).

[527] Zimring, F. E. *The Great American Crime Decline*. Oxford University Press, New York, 2006.

[528] Zou, J., and Schiebinger, L. AI can be sexist and racist – it's time to make it fair. *Nature 559* (2018), 324–326.

[529] Zuckerberg, M. Protecting democracy is an arms race. Here's how Facebook can help. *The Washington Post* (2018).