

DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD
DASH.HARVARD.EDU

HARVARD LIBRARY


Algorithms in Human Decision-Making: A Case Study With the COMPAS Risk Assessment Software

Citation

Vaccaro, Michelle Anna. 2019. Algorithms in Human Decision-Making: A Case Study With the COMPAS Risk Assessment Software. Bachelor's thesis, Harvard College.

Link

https://nrs.harvard.edu/URN-3:HUL.INSTREPOS:37364659

Terms of use

This article was downloaded from Harvard University's DASH repository, and is made available under the terms and conditions applicable to Other Posted Material (LAA), as set forth at

https://harvardwiki.atlassian.net/wiki/external/NGY5NDE4ZjgzNTc5NDQzMGIzZWZhMGFlOWI2M2EwYTg

Accessibility

https://accessibility.huit.harvard.edu/digital-accessibility-policy

Share Your Story

The Harvard community has made this article openly available.
Please share how this access benefits you. Submit a story

Algorithms in Human Decision-Making: A Case Study with the COMPAS Risk Assessment Software

Michelle Vaccaro
Harvard University
Cambridge, MA

Supervisor
Professor Jim Waldo

In partial fulfillment of the requirements for the degree of
Bachelor of Arts in Computer Science
March 29, 2019

Abstract

This thesis uses the COMPAS algorithm as a case study to investigate the role of algorithmic risk assessments in human decision-making. The prior work on the COMPAS algorithm and similar risk assessment instruments focuses on the technical aspects of the tools by presenting methods to improve their accuracy and theorizing frameworks to evaluate the fairness of their predictions. The research does not consider the practical function of the algorithm as a decision-making aid rather than decision-maker.

The first experiment addresses the open question of if algorithmic risk scores impact human predictions of recidivism in a controlled environment with human subjects. The results indicate that the algorithmic risk scores act as anchors that induce a cognitive bias: participants assimilate their predictions to the algorithm's score. In particular, participants who view the low anchor algorithm provide risk scores on average 42.3% lower than participants who view the high anchor algorithm when assessing the same set of defendants. Furthermore, participants can only sometimes perceive when the algorithm's risk scores influence their decisions.

The follow-up experiment explores how the COMPAS algorithm specifically affects the accuracy and fairness of defendant risk assessments. When predicting the risk that a defendant will recidivate, the COMPAS algorithm achieves a significantly higher accuracy rate than the participants who assess defendant profiles (65.0% v 54.2%). Yet when participants incorporate the algorithm's risk assessments into their decisions, their accuracy does not improve. In contrast, when participants view the COMPAS scores the fairness of their predictions changes according to measures of balanced error rates and accuracy equity. They produce more favorable outcomes for white defendants versus black defendants. The experiment also evaluates the effect of presenting an advisement designed to warn of this potential for disparate impact on minorities. The findings suggest, however, that the

advisement does not significantly impact either the accuracy of recidivism predictions or their fairness according to measures of balanced error rates, predictive parity, and accuracy equity.

Based on the theoretical findings from the existing literature, some policymakers and software engineers contend that algorithmic risk assessments like the COMPAS software can alleviate the incarceration epidemic and the occurrence of violent crimes by informing and improving decisions about policing, treatment, and sentencing. But the results from this thesis indicate that if algorithmic risk assessments are to successfully address these problems, future research must investigate their practical role: an input to human decision-makers.

Contents

1 Introduction **1**

1.1 Motivation 1

1.2 Context of the Study 3

1.3 Overview of the Thesis 6

2 Background **7**

2.1 Algorithm-Aided Decisions 7

2.1.1 Anchoring Effect 7

2.1.2 Influence of Risk Assessments on Decision-Making 9

2.2 Algorithm-Aided Decisions and Accuracy 10

2.2.1 Human versus Machine Accuracy 10

2.2.2 Complementary Computing 11

2.3 Algorithm-Aided Decisions and Fairness 12

2.3.1 Fairness through Blindness 13

2.3.2 Fairness through Awareness 14

2.3.3 Inherent Trade-offs 19

2.3.4 Fairness Beyond the Algorithm 20

3 Algorithms as Anchors **22**

3.1 Hypotheses 23

3.2 Task . . . 23
3.3 Participants . . . 27
3.4 Procedure . . . 28
3.5 Quantitative Measures . . . 29
3.5.1 Anchoring Index . . . 29
3.5.2 Deviation . . . 30
3.6 Results . . . 31
3.6.1 Hypothesis 1: Influence . . . 31
3.6.2 Hypothesis 2: Self-Assessment . . . 34
3.6.3 Qualitative Results . . . 34
3.6.4 Participant Experience . . . 36
3.7 Discussion . . . 36
3.7.1 Hypothesis 1: Influence . . . 36
3.7.2 Hypothesis 2: Self-Assessment . . . 37
3.7.3 Limitations and Future Work . . . 38

4 Assessing Accuracy and Fairness 41
4.1 Hypotheses . . . 42
4.2 Task and Procedure . . . 43
4.3 Participants . . . 45
4.4 Quantitative Measures . . . 45
4.5 Results . . . 46
4.5.1 Accuracy . . . 47
4.5.2 Confidence . . . 48
4.5.3 False Positive and False Negative Rates . . . 49
4.5.4 Fairness . . . 50
4.5.5 Error Rate Balance . . . 50
4.5.6 Predictive Parity . . . 52

4.5.7 Overall Accuracy Equity 54
4.5.8 Other Influences on Decision-Making 55
4.5.9 Participant Feedback 56
4.6 Discussion 56
4.6.1 Impact of the Algorithm 58
4.6.2 Impact of the Written Advisement 62
4.6.3 Limitations and Future Work 64

5 Conclusions **69**

A Additional Figures and Tables **72**

References **89**

List of Figures

3.1 An example defendant profile from the experimental survey. . . . 27

3.2 Average recidivism risk score given to defendants in the Low-Score and High-Score treatments. Error bars represent one standard error from the mean. 32

3.3 Average scores participants assigned defendants versus those provided in the defendant profiles in the low-score and high-score treatments. Error bar represent 95% confidence intervals. 33

3.4 Distribution of the anchoring effect among recidivism predictions for defendants. 33

3.5 Relationship between reported influence of and deviation from the risk assessment scores. 34

4.1 Average accuracy of participants in the Control, Score, and Disclaimer treatments. The error bars represent one standard error from the mean. 48

4.2 Participant reported level of confidence in their decisions during the task. The error bars represent one standard error from the mean. 49

4.3 The difference between the false positive rate of participant recidivism predictions for black and white defendants in the control group ($FPR_c^b - FPR_c^w$) and the Score ($FPR_s^b - FPR_s^w$) and Disclaimer ($FPR_d^b - FPR_d^w$) treatment groups. The error bars represent one standard error from the mean. 51
4.4 The difference between the positive predictive value for predictions about black versus white defendants in the Control ($PPV_c^b - PPV_c^w$), the Score ($PPV_s^b - PPV_s^w$) and the Disclaimer ($PPV_d^b - PPV_d^w$) treatment groups. The error bars represent one standard error from the mean. 53
4.5 The difference between the overall accuracy for predictions about black versus white defendants in the Control group ($A_c^b - A_c^w$) and the Score ($A_s^b - A_s^w$) and Disclaimer ($A_d^b - A_d^w$) treatment groups. The error bars represent one standard error from the mean. 55
4.6 Average accuracy of participants in the Control, Score, and Disclaimer treatments in the pilot experiment. The error bars represent one standard error from the mean. 60
4.7 "Person Summary" page of the COMPAS software [1] 66
4.8 Visualization of the defendant risk assessments in the COMPAS software [1] 67

A.1 Survey consent screen. 75
A.2 Survey instructions screen. 76
A.3 Survey pre-questionnaire excerpt. 77
A.4 Survey post-questionnaire excerpt. 78
A.5 Example defendant profile in the Control group. 79
A.6 Example defendant profile in the Score group. 79
A.7 Example defendant profile in the Disclaimer group. 80

List of Tables

2.1 Confusion matrix for the COMPAS algorithm. 14

3.1 Summary of defendant characteristic and algorithm performance in the filtered data set and the experiment's sample. 25
3.2 Summary of answers to the free-response question: *How did you incorporate the COMPAS risk scores into your decisions? (Optional).* 35

4.1 Performance metrics for recidivism predictions of COMPAS algorithm and Control, Score, and Disclaimer treatment groups. . . . 47
4.2 Performance metrics for recidivism predictions by race of the defendant. 50
4.3 Summary of answers to the free-response question: *How did you incorporate the COMPAS risk scores into your decisions? (Optional).* 57

A.1 Participant demographics for experiment one. 73
A.2 Participant demographics for experiment two. 74

Chapter 1

Introduction

1.1 Motivation

The criminal justice system in the United States needs help−the country simultaneously struggles to address the mass incarceration epidemic and reduce occurrences of violent crimes. Since 1970, the American state and federal prison population markedly increased from about 400,000 to 2.2 million today [2]. The United States leads the world in incarceration: approximately 25% of the world's prisoners reside in American prisons, even though the United States only accounts for 5% of the world's population [55]. As such, the United States spends about 75 billion dollars every year on state and local correctional costs, more than other industrialized nations [44]. Despite this large investment, however, the United States faces one of the highest recidivism rates in the world: two in three (68%) individuals released from prison reoffend within five years, and five in six (83%) reoffend within nine years [4]. More than one third (34%) of these rearrests pertain to violent crimes [4] [46].

Predictive algorithms can simultaneously address the problems of mass incarceration and violent crimes by informing and improving decisions about policing,

treatment, and sentencing. In 2006, Camden, New Jersey ranked as one of the most dangerous cities in America with a violent crime rate at 4.4 times the national average [4]. The next year Anne Milgram assumed the role of New Jersey attorney general and attempted to improve public safety by introducing rigorous statistical analysis to the New Jersey police department. Within three years, the murder rate in Camden declined by 41% and the crime rate by 26% [44]. Milgram credits this achievement to the implementation of data-driven models in decision-making processes.

Based on her success in Camden, Milgram argues that similar algorithmic recidivism prediction instruments will improve decisions in the criminal justice system. She explains: "Judges have the best intentions when they make these decisions about risk, but they're making them subjectively...and we know what happens with subjective decision making, which is that we are often wrong" [44]. The algorithmic risk assessment instruments that Milgram supports can recognize individuals likely to commit future violent crimes; furthermore, they can provide people predicted to recidivate with specialized treatment, with the total outcome of reducing crime. The tools can also identify low-risk individuals and administer appropriate non-prison punishments, which would help reduce incarceration rates and costs. Such algorithms therefore possess the potential to both alleviate the incarceration epidemic and prevent future crimes.

Recidivism prediction algorithms can also, importantly, reduce the impact of implicit biases in the criminal justice system and promote more equitable treatment for defendants. A *New York Times* study analyzed 13,876 parole decisions in the state of New York and found that only about 15% of black men were released at their first hearing, compared to about 25% of white men [61]. Researchers also discovered that bail judges demanded on average $7,000 more in bail for black defendants than white defendants in Miami and Philadelphia [7]. Furthermore,

a recent report from the United States Sentencing Commission indicates that black men receive sentences on average 20% longer than similarly situated white men [57]. By introducing increased standardization and constraining potentially biased human discretion, algorithmic risk assessments can reduce these racial disparities.

1.2 Context of the Study

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment software strives to realize these possibilities. The private company Equivant (formerly Northpointe) designed this software, which employs sophisticated quantitative models, broad selections of explanatory theories, extensive ranges of risk and need factors, and integration with criminal justice databases [5]. Concretely, the COMPAS algorithm takes as input responses to 137-questions, some answered by defendants and others derived from criminal records [12]. Using logistic regression, survival analysis, and bootstrap classification methods, the COMPAS model outputs risk scores on a scale from one to ten; these scores indicate the probability of recidivism within two years [12]. Scores from one to four are labeled as Low Risk; five to seven labeled as Medium Risk; and eight to ten labeled as High Risk. States such as Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin use this software in bail, parole, and sentencing decisions [6].

The algorithm may therefore influence the treatment for defendants. For instance, when Judge James Babler of the Wisconsin Circuit Court viewed the COMPAS risk assessment of a defendant classified as high risk, he decided to overturn the plea deal agreed upon by the prosecution and defense to impose a harsher sentence on the defendant. During the appeal of the sentence, however,

the public defender called into question the integrity of the COMPAS score, so Judge Babler reduced the defendant's sentence. He reasoned, "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months" [6].

As this appeal illustrates, ethical and legal objections about the fairness and utility of the COMPAS algorithm quickly arose upon this installation of the COMPAS algorithm into the criminal justice system. These issues garnered national attention in 2016 when ProPublica, an online nonprofit investigative journalism organization, published the article titled *Machine Bias* [6]. This piece demonstrates that the COMPAS algorithm incorrectly classifies black defendants as high risk nearly two times more frequently than white defendants; moreover, the algorithm incorrectly classifies white defendants as low risk nearly two times more frequently than black defendants [6]. Based on these findings, the authors claimed that the COMPAS software demonstrated biases against blacks. ProPublica's allegations sparked heated debates within the computer science community about algorithmic fairness and transparency.

In the 2016 court case *State v. Loomis*, the Wisconsin Supreme Court addressed further challenges against the use of the COMPAS software in sentencing decisions [3]. The defendant alleged that (1) the tool infringed upon due process by unfairly considering gender and (2) the judge's use of the tool violated the right to an individualized sentence. The court rejected the defendant's arguments. In response to the defendant's first claim, the justices found that the inclusion of gender as an input for the algorithm furthered the nondiscriminatory goal of improving the tool's accuracy. Addressing the second argument, the court emphasized that judges retained the discretion to disagree with the algorithm's decisions, so following this logic the defendant still received an appropriately individualized sentence. The court also, however, purposefully cautioned judges

against relying too heavily on algorithmic predictions of recidivism and required a "written advisement" to accompany the risk assessment scores in the presentencing advisement report (PSI) [3]. The justices mandated that this advisement contain five warnings including information about studies, like ProPublica's, that raise questions about the tool's disproportionate classification of minorities as high-risk.

As the *State v. Loomis* decision highlights, the COMPAS software serves as a single and non-determinative factor in human decision-making. Yet much of the prior work assessing the COMPAS software ignores this context and focuses solely on the technical aspects of the algorithm: the tradeoff between its accuracy and various statistical definitions of fairness, for example [14] [36]. While a recent study incorporates human subjects into its experiment, it only compares measures of accuracy and fairness *between* human decisions and COMPAS scores [19]. The literature does not sufficiently investigate the *impact* of these machine predictions on human decision-making. This question, however, is of the utmost importance because human officials serve as the ultimate arbitrators. In the setting of the criminal justice system, human bail officials, parole officers, and judges may ignore the risk scores in certain cases, may defer responsibility to them, or may weigh them differently depending on the race of the defendant. So while the studies concentrating on improving the accuracy and fairness of algorithmic recidivism predictions serve important theoretical purposes, researchers must also examine how humans incorporate these algorithms into their decisions to better understand the practical implications of the algorithms.

1.3 Overview of the Thesis

The COMPAS algorithm can alleviate the problems of mass incarceration and violent crime by improving decisions about parole, bail, and sentencing. As ProPublica's study demonstrates, however, the COMPAS algorithm can also disadvantage minority groups and perpetuate racial biases. To ensure the fair and proper implementation of the instrument, research must examine the algorithm's practical implications. The existing work on the COMPAS algorithm and similar risk assessment instruments focuses on the technical aspects of the tools; it overlooks the practical function of the algorithms as decision-making aids to humans rather than decision-makers themselves. This thesis thus addresses the human aspect of risk assessment algorithms by exploring the following questions: How (if at all) do risk assessment algorithms influence human decisions? And how (if at all) do written advisements accompanying risk assessment algorithms, like that mandated by *State v. Loomis*, affect decision-making?

The next chapter provides the theoretical context to pursue these lines of inquiry by synthesizing prior research on anchoring effects, complementary computing, and algorithmic fairness. Chapter Three presents an experiment inspired by the studies that investigates if algorithmic risk scores impact human predictions of recidivism. Based on the findings from this experiment, Chapter Four describes a follow-up experiment that evaluates how algorithmic risk scores specifically affect the accuracy and fairness of human predictions of recidivism in two cases: with and without the written advisement outlined in *State v. Loomis*. Chapter Five reviews the results from both experiments and contextualizes them within the literature on algorithmic fairness. This chapter also concludes the thesis by underscoring the need for further research on algorithm-assisted decision-making in the criminal justice system and beyond.

Chapter 2

Background

This chapter synthesizes experimental results from psychology, practical examples of human machine collaboration, and technical studies of algorithmic fairness to provide a theoretical basis for the two experiments of the thesis. The work in Section 2.1 on the anchoring effect and econometric studies of risk assessments foreground for the first experiment, and the work in Section 2.2 and 2.3 on the accuracy and fairness of algorithm-aided decisions contextualize the second experiment.

2.1 Algorithm-Aided Decisions

2.1.1 Anchoring Effect

Research in the psychology literature suggests that algorithmic predictions may influence human-decisions through a cognitive bias known as the anchoring effect: when individuals assimilate their estimates to a previously considered standard. Tversky and Kahneman first theorized the anchoring heuristic in 1974 – they published a comprehensive paper that explains the psychological basis of the anchoring effect and provides evidence of the phenomenon through numerous

experiments [56]. In one study, for example, they required participants to spin a roulette wheel that they predetermined to stop at either 10 (low anchor) or 65 (high anchor). After spinning the wheel, participants estimated the percentage of African nations in the United Nations. Tversky and Kahneman found that participants who spun a 10 provided an average guess of 25%, while those who spun a 65 provided an average guess of 45%. They rationalized these results by explaining that people make estimates by starting from an initial value, and their adjustments from this quantity are typically insufficient.

The anchoring index quantifies the effect size of this bias. Let LA denote the value of the low anchor, let HA denote the value of the high anchor, let $\hat{X}_{LA}$ denote the estimates from people exposed to the low anchor, and let $\hat{X}_{HA}$ denote the estimates from people exposed to the high anchor. Then the anchoring index AI can be expressed by Equation 2.1 [11] [32] [31] [40] [56]:

$$AI = \frac{Median(\hat{X}_{HA}) - Median(\hat{X}_{LA})}{HA - LA} \times 100\% \tag{2.1}$$

While initial experiments investigating the anchoring effect recruited amateur participants [56], researchers observe similar anchor effects among experts, too [46]. In their seminal study from 1987, Northcraft and Neale recruited real estate agents to visit a home, review a detailed booklet containing information about the property, and then assess the value of the house [46]. The researchers listed a low asking price in the booklet for one group (low anchor) and a high asking price in the booklet for another group (high anchor). The real estate agents who viewed the high asking price provided valuations 41% greater than those who viewed the lower price, and the anchoring index of the listing price was likewise 41%. Northcraft and Neale conducted an identical experiment among business school students with no real estate experience and observed similar results: the students in the high anchor treatment answered with valuations that exceeded

those in the low anchor treatment by 48%, and the anchoring index of the listing price was also 48%. Their findings, therefore, suggest that anchors such as listing prices bias the decisions of trained professionals and inexperienced individuals similarly.

More recent research finds evidence of the anchoring effect in the criminal justice system. For instance, Englich and Mussweiler conducted a study among judges in which the judges throw a pair of dice and then provide a prison sentence for an individual convicted of shoplifting [21]. The researchers rig the dice so that they land on a low number (low anchor) for half of the participants and high number (high anchor) for the other half. The judges who rolled a low number provided an average sentence of 5 months, whereas the judges who rolled a high number provided an average sentence of 8 months. The difference in responses was statistically significant, and the anchoring index of the dice roll was 67%. Similar studies show that sentencing demands [21], motions to dismiss [33], and damages caps [45] also act as anchors that bias judges' decision-making.

2.1.2 Influence of Risk Assessments on Decision-Making

Strong evidence of the anchoring effect exists in diverse domains including the criminal justice system, yet studies provide conflicting narratives about the role of algorithmic risk assessments as anchors in decision-making. On the one hand, in a study from 2017, Berk investigates the impact of machine learning risk forecasts on parole board decisions in Philadelphia, and he finds that their release decisions did not significantly change after the provision of risk assessments [9]. In a similar study from 2018, Stevenson evaluates algorithmic risk assessments in Kentucky. She finds, however, that after Kentucky passed legislation making pretrial risk assessment mandatory parts of bail decisions: (1) judges granted non-monetary release for low-risk defendants 63% more frequently and (2) judges

granted release for high-risk defendants at approximately the same rate [52]. The different results of these two studies may result from alternative research designs and different geographic settings, among other factors. Importantly, both base their conclusions on econometric analyses of past bail and parole records with some incomplete entries, so they necessarily entail sets of assumptions that may not hold in the real world. Also, neither study incorporates human subjects.

2.2 Algorithm-Aided Decisions and Accuracy

2.2.1 Human versus Machine Accuracy

A recent study from 2018 suggests that algorithms may not outperform humans in the context of recidivism prediction [19]. In this study, Dressel and Farid show that predictions from Amazon Mechanical Turk workers achieved the same levels of accuracy as the COMPAS algorithm. This study conducts two experiments that provide participants with a description of a defendant's alleged crime, sex, age, and criminal history, and asks participants to predict if the defendant would recidivate within two years of their most recent crime. Half of the participants view the races of the defendants and the other half does not. Dressel and Farid find that in both treatments the human accuracy, as measured by proportion of correct positive and negative predictions, does not differ in a statistically significant way from the COMPAS algorithm's accuracy. They also observed that the false positive and false negative rates of participant predictions does not significantly differ from those of the COMPAS algorithm. Dressel and Farid used these results to conclude that the Turk workers were as accurate and fair as COMPAS at predicting recidivism. This study signals an important shift in the literature on algorithmic fairness by incorporating human subjects to contextualize the accuracy and fairness of the algorithms. But their experiment divorces the human

decision-makers and the algorithm when, in fact, they must work in tandem.

2.2.2 Complementary Computing

The field of complementary computing focuses on this suggested coupling of human and algorithmic competencies and demonstrates that humans and machines can achieve better outcomes by working together rather than in isolation [29]. For example, in 1997 IBM's Deep Blue software beat the World Chess Champion Garry Kasparov in a series of six matches [13]. More recent computer programs such as Stockfish and AlphaZero can similarly beat grandmasters [51]. Inspired by his loss, however, Kasparov decided to test the success of Human+AI pairs in chess matches [13]. He found that that the Human+AI team bested the solo human. More surprisingly, he also found that the Human+AI team bested the solo computer, even though the machine outperformed humans. Researchers explain this phenomenon by emphasizing the various dimensions of intelligence [25]. For instance, human chess players excel at long-term chess strategies, but they perform poorly at assessing the millions of possible configurations of pieces. The opposite holds for machines. Due to their different strengths, combining human and machine intelligence produces better outcomes than when each works separately.

The superiority of the human and machine team appears to hold when humans outperform algorithms, too. In the context of a medical image recognition task, medical professionals achieve higher accuracy than the existing machine learning model. When these human experts work in tandem with the computer model, however, their error rates decrease by 85% [49]. Other well-known successful instances of complementarity computing occur in domains such as call routing [29].

Yet the collaborative potential of humans and algorithms lacks rigorous inves-

tigation in the context of risk assessments. In one study, Tan attempts to leverage the human and machine complementarity in this domain [54]. She constructs a hybrid "Human+Machine" model to predict recidivism by combining the predictions from multiple Mechanical Turk workers and the COMPAS scores. Yet this model does not consider the collaboration between humans and machines. The Human+Machine model also does not achieve greater accuracy than either the COMPAS algorithm or the crowd-sourced predictor.

2.3 Algorithm-Aided Decisions and Fairness

The *ProPublica* article Machine Bias initiated the discussion of algorithmic fairness by alleging that the COMPAS software demonstrated biases against black defendants [6]. In this study, the authors analyzed the 2013-2014 arrest records and risk scores of about 7,000 people arrested in Broward County, Florida, and they found that (1) the algorithm incorrectly classified black defendants as high risk nearly two times more frequently than white defendants and (2) the algorithm incorrectly classified white defendants as low risk nearly two times more frequently than black defendants.

This report, however, sparked heated debates within the computer science community about fairness in algorithmic predictions. The creators of the COMPAS software quickly responded and criticized ProPublica's methodologies, namely the classification statistics used to frame their analysis [18]. Their study emphasizes that white and black defendants classified as high-risk recidivate at similar rates, and white and black defendants classified as low-risk also recidivate at similar rates. Additionally, they show that the algorithm achieves equal diagnostic accuracy for white and black defendants across all thresholds of the risk scale. Another study by Flores also contends that the COMPAS algorithm predicts re-

cidivism similarly for black and white defendants, but they base their analysis on logistic regression [23]. They estimate logistic regression models for the relationship between the risk score and outcome and do not find slope or intercept differences for black and white defendants. Fundamentally, however, the COMPAS designers and Flores reach different conclusions from the ProPublica authors because they employ alternative definitions of fairness.

2.3.1 Fairness through Blindness

Recent debates about the statistical definitions of fairness elucidate their embedded social and ethical values. For example, the fairness through blindness approach reflects the anti-classification principle found in legal literature. According to "fairness through blindness," models cannot explicitly consider sensitive attributes such as gender and race [59], which maps onto the anti-classification principle that prohibits classifying individuals "explicitly or surreptitiously" on the basis of protected attributes [8].

Given the high dimensional nature of input data sets, however, the exclusion of race and gender as fields does not necessarily preclude disparate treatment because other features may correlate with them [20]. The COMPAS algorithm, for instance, does not consider race as an input [18]. Features such as income and level of education correlate strongly with race, so different error rates for black and white defendants still occur [6]. Numerous researchers reject input exclusion as a sufficient measure of fairness given the widespread occurrence of similar phenomenon in different data sets [10] [14] [20].

Some academics then propose achieving "blindness" to protected attributes by excluding variables highly correlated with them [59]. The exclusion of these additional factors, however, may decrease the accuracy of the model [20]. In the context of the COMPAS algorithm, according to the lead designer Tim Bren-

	Recidivism Predicted	Recidivism Not Predicted
Recidivism Observed	True Positives (TP)	False Negatives (FN)
Recidivism Not Observed	False Positives (FP)	True Negatives (TN)

Table 2.1: Confusion matrix for the COMPAS algorithm.

nan, "If [income and level of education] are omitted from your risk assessment, accuracy goes down" [6]. Additionally, this feature exclusion approach requires numerous normative decisions, most notably what levels of correlation with protected attributes qualifies as sufficiently "high" to merit removal.

2.3.2 Fairness through Awareness

Most researchers instead advocate for fairness through awareness: explicitly considering sensitive attributes to treat similar individuals similarly [20]. This notion manifests undertones of the anti-subordination theory of fairness as it takes protected attributes like race into account, but for the purpose of preventing disparate impact for historically marginalized groups [8]. Yet when adopting this approach, researchers must also make normative judgements and define both similar individuals and similar treatment. Statistical definitions of fairness reflect the values embedded in these ethical decisions.

Table 2.1 provides a useful framework to explicitly define metrics of algorithmic fairness using a cross-tabulation of the actual outcome by the predicted outcome. The rows and columns represent actual and predicted classifications, respectively. In the context of risk assessment tools like COMPAS, the positive case corresponds to recidivism.

The following section details the most prominent definitions of fairness employed in the last six years [59]. To formally describe these metrics, this section employs the following notation:

- Let $G(x) \in \{b, w\}$ denote the race of defendant x, where b signifies black

and w signifies white.

- Let $Y(x) \in \{0, 1\}$ denote the actual classification result for defendant x, where 0 signifies a negative classification (does not recidivate) and 1 signifies a positive classification (does recidivate).

- Let $S(x) \in \{0, 1, ..., 10\}$ denote the predicted probability of a positive classification for defendant x, where $\forall s \in S$ signifies the value of the defendants' risk scores.

- Let $D(x) \in \{0, 1\}$ denote the algorithm's predicted decision for defendant x, where 0 signifies a negative classification (will not recidivate, low risk) and 1 signifies a positive classification (will recidivate, medium or high risk). The values of D are a function of $S(x)$ as follows:

$$D(x) = \begin{cases} 0 & S(x) < 5 \\ 1 & S(x) \geq 5 \end{cases} \tag{2.2}$$

  Like ProPublica's study, this definition considers medium and high-risk individuals as positive classifications. This grouping of medium and high-risk, notably, has received pushback from the designers of the COMPAS algorithm.

Based on the notation described above, the six most prevalent measures of algorithmic fairness are as follows:

- Statistical parity: An algorithm achieves statistical parity when it predicts that equal percentages of members of protected groups will receive positive and negative classifications [10] [14] [16] [59]. In the context of recidivism, this definition mandates that white and black defendants receive high-risk

scores at the same rate, so the following two equalities must hold:

$$\frac{TP_b + FP_b}{TP_b + TN_b + FP_b + FN_b} = \frac{TP_w + FP_w}{TP_w + TN_w + FP_w + FN_w} \tag{2.3}$$

$$\frac{FN_b + TN_b}{TP_b + TN_b + FP_b + FN_b} = \frac{FN_w + TN_w}{TP_w + TN_w + FP_w + FN_w} \tag{2.4}$$

Statistical parity conditions on the predicted outcome, so Equations 2.2 and 2.3 may also be expressed in probabilistic terms as Equations 2.4 and 2.5, respectively:

$$P(D = 1|G = b) = P(D = 1|G = w) \tag{2.5}$$

$$P(D = 0|G = b) = P(D = 0|G = w) \tag{2.6}$$

Researchers also refer to this measure as equal acceptance rates [10] and group fairness [20]. While much of the machine learning literature employs this measure of algorithmic fairness in problems of employment [39] and admissions [14], people typically do not employ it in the context of recidivism because it may lead to undesirable outcomes: classifying whites as high-risk even if they do not seem likely to recidivate so that equal percentages of whites and blacks are classified as high-risk, for example.

- Error rate balance: An algorithm achieves balanced error rates when the false positive and false negative rates are the same across protected groups [10] [14] [16] [59]. To meet this standard of fairness, the COMPAS algorithm would need to satisfy the two conditions below:

$$FPR_b = \frac{FP_b}{FP_b + TN_b} = \frac{FP_w}{FP_w + TN_w} = FPR_w \tag{2.7}$$

$$FNR_b = \frac{FN_b}{FN_b + TP_b} = \frac{FN_w}{FN_w + TP_w} = FNR_w \tag{2.8}$$

Since error rate balance conditions on both group membership and actual outcomes, equations 2.6 and 2.7 can also be expressed as:

$$P(D = 1|Y = 0, G = b) = P(D = 1|Y = 0, G = w) \tag{2.9}$$

$$P(D = 0|Y = 0, G = b) = P(D = 0|Y = 0, G = w) \tag{2.10}$$

The false positive error rate balance conveys the idea that an algorithm should provide similar predictions for defendants who do not recidivate regardless of race. Likewise, the false negative error rate balance conveys the idea that an algorithm should provide similar predictions for defendants who do recidivate regardless of race.

- Overall accuracy equity: An algorithm achieves overall accuracy equity when the proportion of correct positive and negative predictions is the same among protected groups [10] [14] [16] [59]. The COMPAS algorithm would therefore demonstrate overall accuracy equity if Equation 2.11 holds:

$$\frac{TP_b + TN_b}{TP_b + TN_b + FP_b + FN_b} = \frac{TP_w + TN_w}{TP_w + TN_w + FP_w + FN_w} \tag{2.11}$$

The definition conditions on group membership and predicted outcomes like error rate balance, so Equation 2.11's expression in probabilistic notation appears similar to that of error rate balance:

$$\begin{aligned} &P(Y = 1|D = 1, G = b) + P(Y = 0|D = 0, G = b) = \\ &P(Y = 1|D = 1, G = w) + P(Y = 0|D = 0, G = w) \end{aligned} \tag{2.12}$$

Importantly, however, overall accuracy equity weighs true positive and true negative predictions equally, and it thus assumes they are equally desirable. In the context of the COMPAS algorithm, this assumption equivalently values classifying defendants who recidivate as high risk and classifying defendants who do not recidivate as low risk.

- Predictive parity: An algorithm achieves predictive parity when the probability that a positive prediction for an individual is correct does not depend on any of that individual's protected attributes [10] [14] [16] [59]. In the recidivism prediction setting, the likelihood of recidivism among defendants designated as high-risk should not vary between black and white defendants. As such, the following equation must hold:

$$\frac{TP_b}{TP_b + FP_b} = \frac{TP_w}{TP_w + FP_w} \tag{2.13}$$

  Researchers refer to the ratio $\frac{TP}{TP+FP}$ as the positive predictive value (PPV) of the algorithm [10] [14] [16] [59]. The metric conditions on group membership and predicted outcomes, so Equation 2.14 can also be written as:

$$P(Y = 1|D = 1, G = b) = P(Y = 1|D = 1, G = w) \tag{2.14}$$

  Predictive parity conveys the value that the accuracy of high-risk classifications should be the same for both black and white defendants.

- Calibration: An algorithm achieves calibration when the proportion of people who experience the predicted outcome does not vary across protected groups for each predicted value [10] [14] [16] [59]. This definition mirrors that of predictive parity, but considers the PPV for each value of s. Put

formally:

$$\forall s \in S, P(Y = 1|S = s, G = b) = P(Y = 1|S = s, G = w) \tag{2.15}$$

For the COMPAS algorithm, calibration would occur if, for every given risk score, the percentage of black defendants who recidivate equals the percentage of white defendants who recidivate.

2.3.3 Inherent Trade-offs

The measures of fairness described above are intrinsically related: they rely on the cell counts TP, TN, FP, and FN from the confusion matrix in Table 1, and inherent trade-offs exist between them. Namely, when base rates differ across protected groups, an algorithm cannot satisfy predictive parity and error rate balance simultaneously [15] [16] [36]. Expressing the false positive rate (FPR) and false negative rate (FNR) in terms of the prevalence (p) of a given trait illustrates this result:

$$\begin{aligned} FPR &= \frac{FP}{FP + TN} \\ &= \frac{TP + FN}{FP + TN} \times \frac{\frac{FP}{TP+FP}}{\frac{TP}{TP+FP}} \times \frac{TP}{TP + FN} \\ &= \frac{p}{1 - p} \times \frac{1 - PPV}{PPV} \times (1 - FNR) \end{aligned}$$

As the equations above demonstrate, if $FPR_b = FPR_w$ and $FNR_b = FNR_w$ (balanced error rates) and $p_b \neq p_w$, then PPV_b cannot equal PPV_w (predictive parity). Similarly, if $PPV_b = PPV_w$ and $p_b \neq p_w$, then FPR_b cannot equal FPR_w or FNR_b cannot equal FNR_w.

Since white and black defendants recidivate at different rates, no risk predic-

tion algorithm can achieve predictive parity (COMPAS's measure of fairness) and equal false positive and false negative rates (ProPublica's measure of fairness). So while several researchers have successfully devised processes to alter algorithms to achieve certain definitions of fairness [22] [28] [35], their methods necessarily prioritize selected values and make normative judgements.

2.3.4 Fairness Beyond the Algorithm

Given that any machine learning predictor like the COMPAS algorithm will be unfair according to some metric, the Wisconsin Supreme Court wrote an advisement about the COMPAS algorithm's potentially disparate impact on minorities [3]. The Court mandated that this disclaimer accompany the COMPAS risk scores in their presentation to judges. The written advisement, moreover, presents a novel and non-technical approach to algorithmic fairness by looking beyond the algorithm's input data, inner structure, and output predictions. But the provision of the advisement highlights an open question within complementary computing: namely, how (if at all) does such information impact the algorithm's influence on human decision-making? On the one hand, these advisements can promote greater understanding of the abilities and limitations of the algorithm and thus help users work more effectively with it. On the other hand, research of disclaimers in advertising contexts suggests they may have the opposite effect as intended. For example, Green and Armstrong find evidence that mandatory government disclaimers increase confusion for consumers or are otherwise harmful [27]. But little, if any, research exists about disclaimers for algorithms and their effects on human decision-making.

Chouldechova's recent paper about algorithm-assisted decision making in child maltreatment hotline screening decisions highlights the need to consider the human aspect of algorithmic fairness, particularly in contexts when humans serve

as the final decision arbitrators [15]. In the study, Chouldechova constructs a risk prediction model using government administrative data to target services to the riskiest cases of child abuse and maltreatment. Since such an instrument could disadvantage some communities – for example certain racial groups – the study analyzes the fairness of the algorithm according to its predictive parity, accuracy equity, and error rates. During the implementation of the tool, however, almost 25% of supervisors override mandatory screen-ins. Since humans selectively consider the scores, Chouldechova highlights that "the fairness and accuracy properties of the tool will not carry over into equitable and effective decision-making" [15]. As such, while the analysis of the fairness of the risk-assessment provides valuable insights, it does not address the accuracy or fairness of the algorithm's practical outcomes as a decision-making aid rather than decision-maker.

The next two chapters present experiments influenced by the questions, methods, and findings of these studies. The experiment in Chapter 3 draws particular inspiration from the work of Dressel [19], Kahneman [34], and Northcraft [46]: the methodology mirrors that of Dressel and Northcraft's, the quantitative measures of the anchoring effect come from Kahneman, and the hypotheses rely on results from Northcraft and Kahneman. The experiment in Chapter 4 similarly leverages the insights from Dressel [19], Choulechova [14], and Case [13]: the methodology of this experiment also echoes that of Dressel's, the quantitative measures of accuracy and fairness reflect those in Choulechova's paper about algorithmic fairness, and the hypotheses stem from the work of Case in complementary computing.

Chapter 3

Algorithms as Anchors

As prior work suggests that algorithmic risk assessments demonstrate greater accuracy than judges [37], some policymakers contend that federal and state courts should employ these tools in decisions about sentencing, bail, and parole to improve their accuracy [44]. Yet others highlight issues of racial biases embedded in such algorithms, and they allege that the introduction of these tools into the criminal justice system will exacerbate racial inequalities [6]. While these claims reach opposing conclusions, they both rest on the presumption that the COMPAS risk scores ultimately influence human decisions. This experiment seeks to investigate that assumption by gathering quantitative and qualitative measures of the impact of algorithmic risk scores on human predictions of recidivism.

Participants complete an online survey requiring them to synthesize information about defendants and report on the defendants' risk of recidivism. The experiment entails a 1x2 between-subjects design where the two treatments are as follows: Low-Score, in which participants view the defendant profile accompanied by a low risk score; and High-Score, in which participants view the defendant profile accompanied by a high risk score. The experiment thus investigates the following two questions in the context of recidivism risk assessments: (1) Do

algorithmic risk scores affect human predictions of recidivism? (2) Do people recognize if algorithmic risk scores influence their predictions of recidivism?

3.1 Hypotheses

Given these guiding questions, the research hypotheses for the experiment are as follows:

1. Hypothesis 1 (Influence): Prior studies in psychology indicate that people who view low anchors provide lower estimates than those who view high anchors in domains such as real estate [46], meteorology [34], and sentencing [21], so this study hypothesizes that participants in the Low-Score treatment will assign lower risk scores to defendants than participants in the High-Score treatment.

2. Hypothesis 1 (Self-Assessment): Prior research indicates that individuals do not perceive when an anchor biases their decision-making [46], so this study hypothesizes that participants will not accurately self-assess the influence of the algorithm on their decisions.

3.2 Task

The experimental task entails evaluating a sequence of 40 defendant profiles that convey information about the defendants' genders, races, ages, criminal charges, and criminal histories. These profiles describe real people arrested in Broward County, Florida, based on information from the data set that ProPublica used in their analysis [6]. While this dataset originally contains 7214 entries, this study applies the following filters before sampling the 40 profiles that it presents to participants.

- Limit to Black and White Defendants: Prior work on the accuracy and fairness of the COMPAS algorithm limits their analysis to white and black defendants [14] [16] [19]. To compare the results from this experiment to those in prior studies, this study only considers the subset of defendants who identify as either African-American (black) or Caucasian (white).

- Exclude Cannabis Crimes: Additionally, during a pilot study, participants indicated confusion over cannabis related crimes such as "Possession of Cannabis," "Purchase of Cannabis," and "Delivery of Cannabis." In the free response section of the survey, they noted comments such as: "cannabis is fully legal here." To avoid confusion about the legality of cannabis in various states, this study also excludes defendants charged with crimes containing the term "cannabis." Applying these filters removes 1064 (14.7%) of the entries in the data set. The restriction of these entries, however, does not significantly affect the observed accuracy (as measured by overall accuracy) or fairness (as measured by predictive parity, equalized odds, and balanced error rates) of COMPAS algorithm on defendants in the dataset.

From this filtered data set, this study randomly samples 40 defendants to present in the experiment. As shown in Table 3.1, the descriptive statistics of the sample resembles that of the population. The observed accuracy or fairness of COMPAS algorithm on defendants in the dataset and the sample does not significantly vary, either.

For each of the 40 defendants, this study generates a profile that contains information about their demographics, alleged crime, criminal history, and algorithmic risk assessment. Each descriptive paragraph assumes the following format, which builds upon that used in Dressel and Farid's study [19]:

```
The defendant is a [RACE] [SEX] aged [AGE]. They have been
```

	Pop (Black) N=3696	Pop (White) N=2454	Pop (Overall) N=6150	Sample (Black) N=27	Sample (White) N=13	Sample (Overall) N=40
Defendant Profile						
Male	82.35%	76.89%	80.18%	77.78%	61.54%	72.5%
Female	17.65%	23.11%	19.82%	22.22%	38.46%	27.5%
Average Age	32.74	37.73	34.73	31.93	41.00	34.88
Felony	68.91%	60.31%	65.48%	66.67%	61.54%	65.00%
Misdemeanor	31.09%	39.69%	34.52%	33.33%	38.46%	35.00%
# of Priors	4.44	2.59	3.70	5.22	2.08	4.20
Juv Fel (Avg)	0.10	0.03	0.07	0.11	0.00	0.08
Juv Mis (Avg)	0.14	0.04	0.10	0.15	0.00	0.10
2y recidivism	51.43%	39.36%	46.62%	55.56%	38.46%	50.00%
Algorithm Performance						
Avg Score	5.37	3.74	4.72	4.93	3.08	4.33
Low Risk	41.18%	65.20%	50.76%	44.44%	76.9%	55.00%
Medium Risk	31.09%	23.55%	28.08%	44.44%	7.69%	32.50%
High Risk	27.73%	11.25%	21.15%	11.11%	15.38%	12.50%
Accuracy	63.72%	67.06%	65.09%	64.75%	65.38%	64.75%
FNR	30.00%	50.12%	36.88%	28.05%	50.00%	34.21%
FPR	43.17%	21.93%	33.18%	50.00%	23.91%	36.05%
PPV	63.97%	59.32%	62.64%	74.68%	59.26%	70.75%

Table 3.1: Summary of defendant characteristic and algorithm performance in the filtered data set and the experiment's sample.

```
charged with: [CRIME CHARGE]. This crime is classified as
a [CRIMINAL DEGREE]. They have been convicted of [NON-JUVENILE
PRIOR COUNT] prior crimes. They have [JUVENILE-FELONY COUNT]
juvenile felony charges and [JUVENILE-MISDEMEANOR COUNT] juvenile
misdemeanor charges on their record. COMPAS is risk assessment
software that uses machine learning to predict whether a defendant
will commit a crime within the next two years. The COMPAS
risk score for this defendant is [SCORE NUMBER]: [SCORE LEVEL].
```

The Low-Score and High-Score treatments assign risk scores based on the original COMPAS score according to the following formulas:

$$\text{Low-Score} = \min(0, COMPAS - 3) \tag{3.1}$$

$$\text{High-Score} = \max(10, COMPAS + 3) \tag{3.2}$$

Essentially, Low-Score subtracts three points from the original COMPAS score, and High-Score adds three points to the original COMPAS score.

Upon seeing each profile, participants must provide their own risk assessment scores for the defendant and indicate if they believe the defendant will recidivate within two years. Using a dropdown menu, they answer the followings questions as illustrated in Figure 3.1.

For the later question, participants must provide a "Yes" or "No" response, which permits a more thorough analysis of the accuracy and fairness of the recidivism predictions. The ProPublica analysis received much criticism for considering low risk scores as negative predictions of recidivism and considering both medium and high risk scores as positive predictions of recidivism [18] [23]. This second question provides a more concrete measure of negative and positive predictions, so

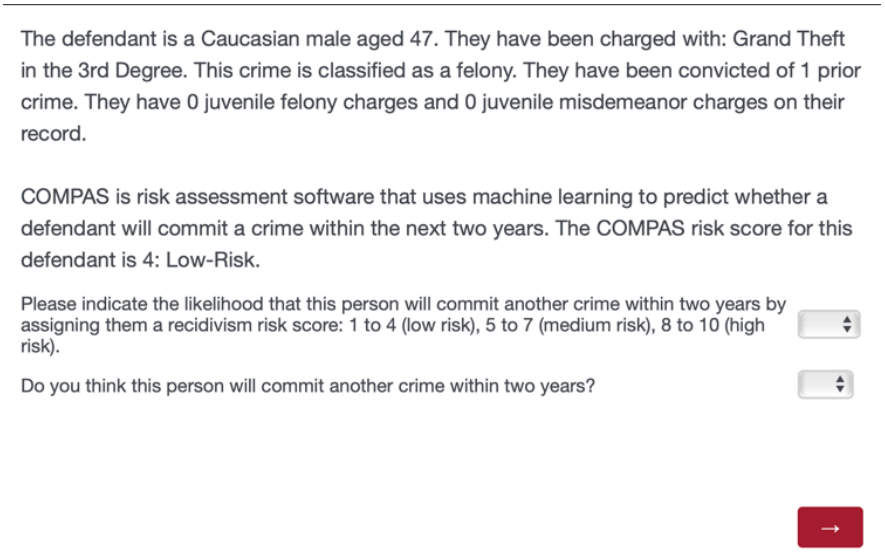


Figure 3.1: An example defendant profile from the experimental survey.

this study does not need to render an ordinal variable with ten different rankings as binary.

Both questions, moreover, require participants to synthesize the information in each defendant profile and assess the defendant's risk of recidivism. Each treatment shows the same 40 defendants, but in a randomized order to control for priming and learning biases. Participants view defendant profiles as discrete task blocks, so they cannot go back and alter their previous answers later in the survey.

3.3 Participants

The study deploys remotely through the Qualtrics platform, and it recruits 100 respondents through Amazon Mechanical Turk. All workers can view the task title "Predicting Crime," task description "Answer a survey about predicting crime," and the key words associated with the task "survey, research, and criminal justice." Only workers living in the United States, however, can complete the

task, and they can do so only once. During a pilot of the survey among an initial test group of five individuals, the survey required an average of 15 minutes to complete. As the length and content of the survey resemble that of Dressel and Farid's [19], this study adopts their payment scheme and pays workers $1.00 for completing the task and provides a $2.00 bonus if the overall accuracy of the respondent's predictions exceeds 65%. This payment structure incentivizes participants to pay close attention and provide their best responses throughout the task [19] [50].

3.4 Procedure

The experimental survey contains five sections: (1) Consent form; (2) Pre Questionnaire; (3) Instructions; (4) Risk Assessments; (5) Post Questionnaire. First, participants must read an information page that describes basic information about the expected survey length, the right to withdraw at any time, the minimum age requirement of 18 years old, and the data collection practices. Participants must indicate that they consent to participate in the study to proceed to the next page. Participants then answer an initial questionnaire collecting demographic information including gender, age, ethnicity, educational attainment. They also answer two questions assessing their familiarity with the United States criminal justice system and machine learning on five-point Likert scales.

Next, participants read an instructions page providing them with the following information about their task:

> `In this task, you will evaluate information about an individual charged with a criminal offense. The data you will view pertains to crimes committed in the United States between 2013 and 2014. You will see the defendant's age, race, gender, previous criminal`

> history, and COMPAS risk score. COMPAS is risk assessment software that uses machine learning to predict whether a defendant will commit a crime within the next two years. After evaluating the defendant profile, you will (1) assign the person a risk score and (2) predict if the person will commit another crime within two years. You will earn $1.00 for completing the study. If the overall accuracy rate of your predictions exceeds 65%, then you will receive an additional $2.00 bonus reward. As such, please answer the following questions to the best of your ability.

After completing the pre-experiment phase, participants complete the task: they view the defendant profiles and provide risk assessments for each individual. Participants view one defendant description at a time, and they cannot go back to previous profiles and change their answers. They see these profiles in a randomized order. Upon finishing the task block, they answer a final series of questions about their perception of the COMPAS algorithm, their confidence in their decisions, and their satisfaction with the study on seven-point Likert scales. In this section, participants also have the option to provide further information about their decision-making process and their general experience with the study via free response. Figures A.1 - A.4 in the Appendix provide screenshots of the pages in each of these sections.

3.5 Quantitative Measures

3.5.1 Anchoring Index

The anchoring index quantifies the effect size of the psychological impact of anchors on decision-making. Let LA denote the value of the low anchor, let HA

denote the value of the high anchor, let $\hat{X}_{LA}$ denote the estimates from people exposed to the low anchor, and let $\hat{X}_{HA}$ denote the estimates from people exposed to the high anchor. Then the anchoring index AI can be expressed by Equation 3.1 [11] [31] [56]:

$$AI = \frac{Median(\hat{X}_{HA}) - Median(\hat{X}_{LA})}{HA - LA} \times 100\% \tag{3.3}$$

For example, if the algorithm assigns defendant one a risk score of 7 in the High-Score treatment, then 7 serves as the high anchor so $HA = 7$. Likewise if the algorithm assigns defendant one a risk score of 3 in the Low-Score treatment, then 3 serves as the low anchor so $LA = 7$. If participants in the High-Score treatment assign defendant one a median risk-score of 5, then $Median(\hat{X}_{HA}) = 5$. Similarly, if participants in the Low-Score treatment assign defendant one a median risk-score of 2, then $Median(\hat{X}_{HA} = 2)$. In this case, the anchoring index for defendant one is: $AI_1 = \frac{5-2}{7-1} \times 100\% = 50\%$. An anchoring index of 100% indicates that people entirely adopt the anchor, and an anchoring index of 0% indicates that people completely ignore the anchor.

3.5.2 Deviation

To identify instances in which participants perceive the algorithm's risk score as an overestimate or underestimate, this study measure participants' deviation from the algorithm's risk score: the difference between the score they assign and the score they view [26]. To provide a formal definition, let p_i^j denote the score participant j assigns to defendant i, and let a_i denote the score the algorithm assigns to defendant i. Then participant j's deviation from the algorithm for defendant i can be expressed as:

$$d_i^j = p_i^j - a_i \tag{3.4}$$

Values of $d < 0$ indicate that a participant believes the algorithm's risk score overestimated the defendant's risk of recidivism. Values of $d = 0$ indicate that a participant agrees with the algorithm's risk score. And values of $d > 0$ indicate that a participant believes the algorithm's risk score underestimated the defendant's risk of recidivism.

3.6 Results

The 100 Mechanical Turk workers completed the experiment over the course of two days in February 2019. Of these participants, 44% identify as female, 55% identify as male, and 1% do not identify with a binary gender. The vast majority identify as white (86%) with only 2% identifying as African-American. The majority (56%) list college as their highest level of education with 23% reporting high school as their highest level of education and 21% reporting master's degrees. When asked to rank their familiarity with the criminal justice system on a 5-point Likert scale, the majority of participants report 3, which maps to "somewhat familiar". The median result is also 3, and mean result is 3.11. Participants report less familiarity with machine learning. The majority of participants respond "not very familiar," which corresponds to a value of 2. As such, the median result is 3, but the mean result is 2.77. Table A.1 summarizes these participant demographics.

3.6.1 Hypothesis 1: Influence

The scores that participants assigned defendants highly correlates with those that they viewed in the defendants' profile descriptions. As such, participants in the Low-Score treatment provide risk scores on average 42.3% lower than participants in the High-Score treatment when assessing the same set of defendants (Figure

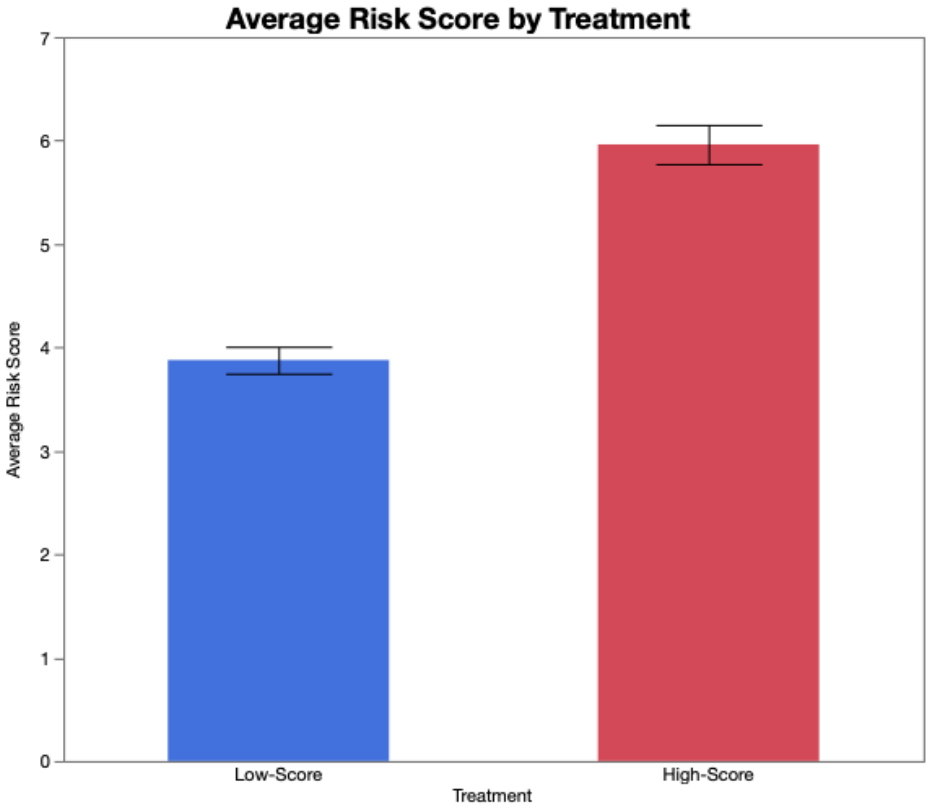


Figure 3.2: Average recidivism risk score given to defendants in the Low-Score and High-Score treatments. Error bars represent one standard error from the mean.

3.2). The average risk score from respondents in the Low treatment is 3.88 (95% CI 3.39-4.36) while the average risk score from respondents in the High treatment is 5.96 (95% CI 5.36-6.56). A two-sided t-test reveals that this difference is statistically significant ($p < 0.0001$).

Put in other words, the risk assessment scores that participants view in the defendant profiles serves as anchors that impact participants' decision-making. As shown in Figure 3.4, the anchoring indices for predictions about the defendants cluster between 50% and 70%. The average anchoring index across all 40 defendants is 56.71% (95% CI 52.97%-60.44%).

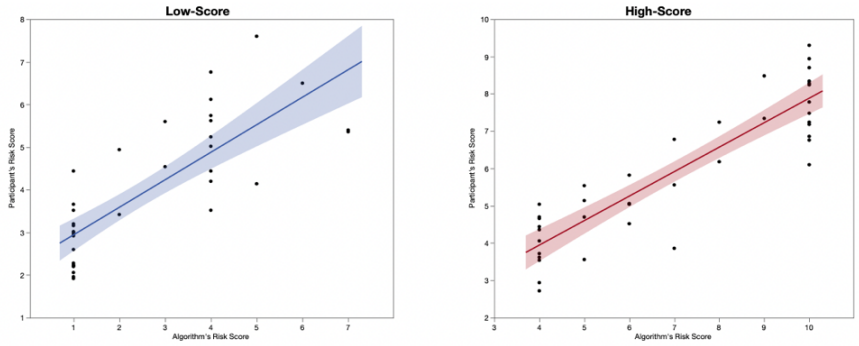


Figure 3.3: Average scores participants assigned defendants versus those provided in the defendant profiles in the low-score and high-score treatments. Error bar represent 95% confidence intervals.

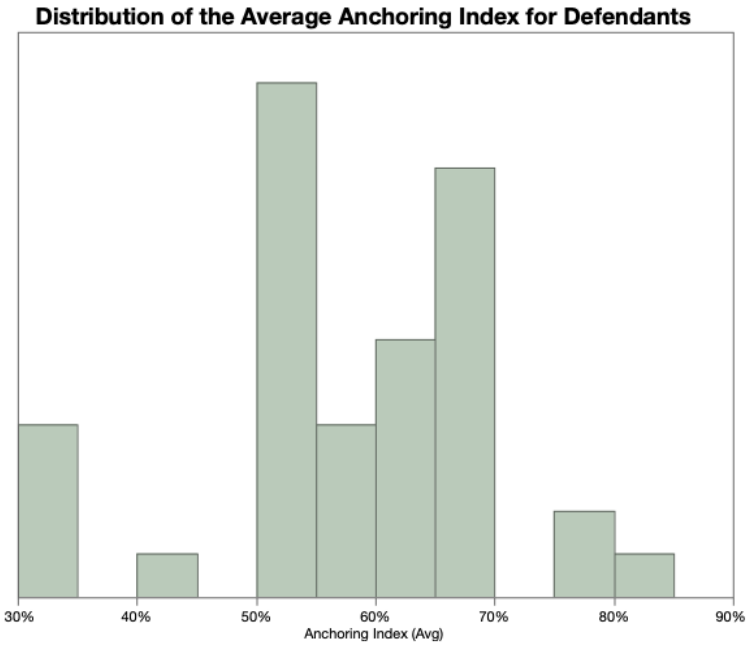


Figure 3.4: Distribution of the anchoring effect among recidivism predictions for defendants.

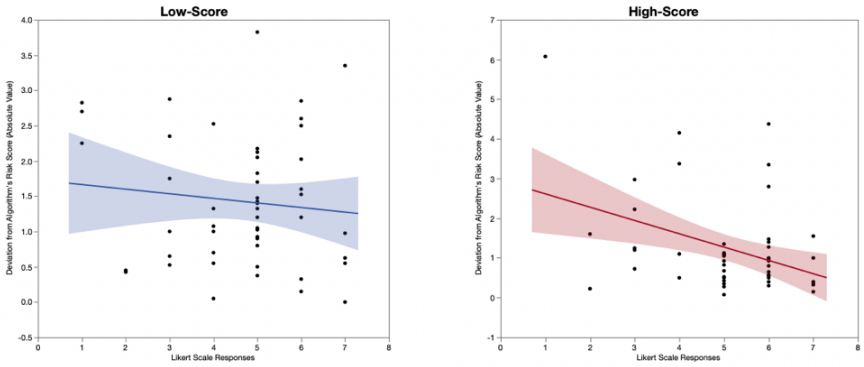


Figure 3.5: Relationship between reported influence of and deviation from the risk assessment scores.

3.6.2 Hypothesis 2: Self-Assessment

One question in the exit survey of the experiment requires participants to self-assess their predictions. They must indicate their agreement with the following statement on a 7-point Likert scale: "The COMPAS risk score influenced my decisions." This measure of perceived influence significantly correlates with the actual influence of algorithm on participants' predictions in the High-Score treatment ($p = 0.0051$) but not the Low-Score treatment ($p = 0.4453$) (Figure 3.5).

3.6.3 Qualitative Results

When reflecting on the role of the COMPAS algorithm in their decision-making process at the end of the survey, participants indicate common themes including using the algorithm's score as a starting point and as a verification of their own decisions. Table 3.2 summarizes these participant comments by their treatment group and role of the algorithm in their decision-making.

Algorithm Role	Low Treatment	High Treatment
Factor	"I took them into consideration but still made my own decisions."	"I took it into account, but did not count on it 100 percent." "I used it as ONE factor."
Tipping Point	"I used it if I was wavering on a score." "I would look at it if I was close on which way to go."	"For those cases where I felt a 50/50 chance, I sided with the Compas score."
Validation	"Only considered it when it seemed to coincide with my own judgment."	"I looked to see if it was similar to what I thought."
Guideline	"I used it to target my general range of scores, unless I had reason to strongly disagree." "I used the scores to base the start value of my score, read the description of their crime and modified the score."	"I kind of started with the Compas risk score, and then raised or lowered the score based on previous criminal history (or lack of one)." "I used it as a guideline to structure my decisions on."
Deference	"I always considered it as near perfect"	NA
Ignored	"I usually ignored it because it didn't seem like it made much sense to me, but who knows." "It didn't seem very consistent or accurate so I didn't factor it in much, if at all."	"I HATE rubrics. You are looking at people, dynamic people, and a computerized rubric or other type of system designed to assess risks is completely ignoring so many other *very important* circumstances that may affect these odds."

Table 3.2: Summary of answers to the free-response question: *How did you incorporate the COMPAS risk scores into your decisions? (Optional).*

3.6.4 Participant Experience

When providing their general feedback about the study, participants give primarily positive responses like "Thanks," "Very engaging," and "Great HIT!" At the end of the study, 98% of participants indicate that they found the instructions for the task clear, and 91% of participants indicate satisfaction with the task.

3.7 Discussion

This study demonstrates that algorithmic risk assessments significantly impact human predictions of recidivism. Participants in the Low-Score treatment assign significantly lower risk scores than those in the High-Score treatment, even though they assess the same group of defendants. As such, the algorithm's scores act as anchors that exhibit an average anchoring index of 56.71% in the experiment. Participants can only sometimes perceive when the algorithm's risk scores influence their decisions. This section revisits the initial research hypotheses and discusses the connections between the experiment's results and the integration of recidivism prediction algorithms into the criminal justice system.

3.7.1 Hypothesis 1: Influence

The results from this study indicate that algorithmic risk predictions serve as anchors that bias human decision-making. Participants in the Low-Score treatment provide an average risk score of 3.88, while participants in the High-Score treatment assign an average risk score of 5.96. The average anchoring index across all 40 defendants is 56.71%. This anchor measure mirrors that found in prior psychology literature [24] [34] [46]. For example, one study investigates the anchoring bias in estimations by asking participants to guess the height of the tallest redwood tree [34]. The researchers provide one group with a low anchor

of 180 feet and another group with a high anchor of 1,200 feet, and they observe anchoring index of 55%. Scholars observe similar values of the anchoring index in contexts such as probability estimates [56], purchasing decisions [60], and sales forecasting [17].

Even though this type of cognitive bias occurs among participants with little training in the criminal justice system, prior work suggests that the anchoring effect varies little between non-experts and experts in a given field. For instance, Northcraft and Neale conduct an experiment in which real estate agents visit a house, review a booklet containing information about the property, and then assess the value of the house. The researchers include a low asking price in the booklet for one group (low anchor) and a high asking price in the booklet for another group (high anchor), and they find an anchoring effect of 41%. The researchers conducted the same study with business school students with no real-estate experience and find a similar anchoring index of 48%. This study thus suggests that the anchoring effect of algorithmic risk assessments among judges, bail, and parole officers would mirror that of the participants in this experiment. Numerous prior studies demonstrate that these officials are susceptible to forms of cognitive bias like anchoring, too [21] [45]. On the other hand, prior work indicates that experts rely less heavily on other's advice than non-experts [30] [41], so officials in the criminal justice system may opt to ignore an algorithm's risk score in lieu of their own judgements about defendants. To better determine the influence of these risk assessment algorithms, future work should employ judges, bail officers, and parole officials as participants.

3.7.2 Hypothesis 2: Self-Assessment

As prior studies of the anchoring effect observe [21] [34] [56], the participants in the Low-Score treatment do not understand the degree to which the algorithm's

risk scores influence their decisions. For example, in Northcraft and Neale's study about the anchoring effect of home asking prices, participants insisted that the asking price of the house did not affect their judgements [46]. Likewise, in the Low-Score treatment, participants' perceived influence of the algorithm's risk scores on their predictions does not correlate with their adherence to it. In the High-Score treatment, in contrast, a statistically significant relationship does exist between the perceived influence of the algorithm and the adherence to it.

This result may occur because people are more likely to believe an artificially high-risk score than an artificially low one. In fact, in the optional comments section at the end of the study, some participants explicitly mentioned that they perceived the scores as too low. Responses to this end include: "Seemed like COMPAS was far too lenient" and "I really didn't understand why so many of the individuals were repeatedly labelled as low-risk." Participants in the high-anchor treatment, in contrast, did not comment on the harshness of the scores nor the number of individuals labelled as high-risk.

3.7.3 Limitations and Future Work

The participants in this experiment consist of workers recruited through the Amazon Mechanical Turk platform. To ensure manageable task length, each participant only evaluates 40 defendant profiles randomly sampled from the arrest records in Broward County, Florida. To measure the anchoring effect and influence of the algorithm's risk scores on participant predictions, this set of defendants remains constant across all participants in both the Low-Score and High-Score treatments. While the accuracy and gender ratio of this sample resemble that of the population, the sample includes only a small representation of the various crime convictions and paints a limited picture of the various criminal backgrounds (i.e. juvenile criminal history, number of prior convictions, etc). To

increase external validity, future work should conduct similar studies at a greater scale by employing a larger sample of defendants. This work could also investigate the relationship between different defendant histories and the influence of recidivism prediction instruments on human assessments of their risk to reoffend.

While hosting this study on Mechanical Turk buids upon prior work in algorithmic fairness that also recruits participants through this platform [19] [26] [54], it also introduces questions about participant accuracy, diversity, and credibility. Because participants automatically receive $1.00 for completing the task, some may not pay attention to their accuracy and simply try to finish the experiment as soon as possible. The $2.00 bonus provides a monetary incentive for participants to provide their best effort throughout the task and attempts to ensure valid responses. The comments workers left in the feedback section inquire about whether they earned the bonus, so they indicate that this potential monetary reward did motivate the workers. Yet some may still disregard this incentive and thus skew the results. Subsequent studies could lower or eliminate the base payment while increasing the bonus reward and evaluate if results differ when potential for reward is larger.

Additionally, much like the Amazon Mechanical Turk workforce, the participants in this study lack racial diversity. Of the 100 workers recruited for the study 86% identify as white and only 2% identify as black. Given the small number of non-white participants, this study could not measure if race affects predictions of recidivism. Yet prior work indicates that individuals of different races harbor different implicit racial biases [42]. Future research, therefore, could specifically recruit equal numbers of white and black participants through Amazon Mechanical Turk and test if they incorporate algorithmic risk scores in their predictions differently.

Despite these limitations, the results of this experiment provide convincing

evidence that algorithmic risk assessments exert influence on human decision-making. The anchoring index of the algorithm's risk score, in particular, provide quantitative evidence that algorithmic risk scores influence predictions of recidivism, and participant free responses provide qualitative evidence that further supports this conclusion. But while they answer the question about if algorithmic risk assessments affect human decision-making, they do not answer the question about how, when or why. Such questions merit further investigation, particularly in light of the opposing claims that contend (1) these algorithms will improve the accuracy of decisions in the criminal justice system and (2) these algorithms will lead to unfair decisions in the criminal justice system. The next chapter presents a follow-up study that pursues these lines of inquiry.

Chapter 4

Assessing Accuracy and Fairness

The first experiment demonstrates that algorithmic risk assessments serve as anchors that induce a cognitive bias and thus exert influence on human decision-making. The second experiment builds upon these results by exploring the open question of *how* COMPAS scores specifically affect the accuracy and fairness of human risk assessments. In the 2016 court case *State v. Loomis*, the Wisconsin Supreme Court expressed concern about the role of COMPAS scores in sentencing decisions, so they mandated the inclusion of a list of advisements accompanying the presentation of the scores [3]. Given this recent development, experiment two also investigates the effect of the written advisement on human predictions of recidivism.

Much like the first experiment, this experiment entails an online survey that requires participants to synthesize information about defendants and report on the defendants' risk of recidivism. In this study, however, the experiment entails a 1×3 between-subjects design with the following treatments: Control, in which participants only see the defendant profiles; Score, in which participants see the defendant profiles and the defendant COMPAS scores; and Disclaimer, in which participants see the defendant profiles, the defendant COMPAS scores, and the

written advisement about the COMPAS algorithm. This design allows an examination of the following two questions: (1) Does providing COMPAS scores to humans affect the accuracy or fairness of their predictions of recidivism? (2) Does providing a written advisement with the COMPAS score affect the influence of the score on human predictions of recidivism?

4.1 Hypotheses

Given prior work in complementary computing that shows coupling machine and human predictions can improve outcomes [25] [29] [49], this study forms the following research hypotheses to address question one:

- Hypothesis 1.1 (Accuracy): Individuals who view the COMPAS scores will predict recidivism with greater accuracy that those who do not see the scores.

- Hypothesis 1.2 (Confidence): Individuals who view the COMPAS scores will indicate higher confidence in their predictions than those who do not see the scores.

Additionally, since previous studies demonstrate that the COMPAS algorithm exhibits high false positive rates for black defendants and high false negative rates for white defendants [6], this study also forms the following research hypotheses to address question one:

- Hypothesis 1.3 (False Positive Rate): Individuals who view the COMPAS scores will falsely predict that black defendants will recidivate at a higher rate than those who do not see the scores.

- Hypothesis 1.4 (False Negative Rate): Individuals who view the COMPAS

scores will falsely predict that white defendants will *not* recidivate at a higher rate than those who do not see the scores.

Next, because psychology research demonstrates that disclaimers do not significantly impact decision-making [27], this study forms the following research hypotheses to address question two:

1. Hypothesis 2.1 (Accuracy): Individuals who view the COMPAS scores with the accompanying written advisement will predict recidivism with the same accuracy as those who view only the COMPAS score.

2. Hypothesis 2.2 (Confidence): Individuals who view the COMPAS scores with the accompanying written advisement will indicate the same confidence in their predictions as those who view only the COMPAS score.

3. Hypothesis 2.3 (False Positive Rate): Individuals who view the COMPAS scores with the accompanying written advisement will falsely predict that black defendants will recidivate at the same rate as individuals who view only the COMPAS score.

4. Hypothesis 2.4 (False Negative Rate): Individuals who view the COMPAS scores with the accompanying written advisement will falsely predict that white defendants will not recidivate at the same rate as individuals who view only the COMPAS score.

4.2 Task and Procedure

The experimental task entails evaluating a sequence of the same 40 defendants presented in experiment one. Like the previous experiment, each treatment shows the identical set of defendants, but in a randomized order to control for priming

and learning biases. Participants still view defendant profiles as discrete task blocks, so they cannot go back and alter their previous answers later in the survey. The content of the profiles, however, differs from that in experiment one as follows:

The descriptive paragraph in the Control condition reads:

> `The defendant is a [RACE] [SEX] aged [AGE]. They have been charged with: [CRIME CHARGE]. This crime is classified as a [CRIMINAL DEGREE]. They have been convicted of [NON-JUVENILE PRIOR COUNT] prior crimes. They have [JUVENILE-FELONY COUNT] juvenile felony charges and [JUVENILE-MISD COUNT] juvenile misdemeanor charges on their record.`

The descriptive paragraph in the Score treatment adds the following information in the defendant profile:

> `COMPAS is risk assessment software that uses machine learning to predict whether a defendant will commit a crime within the next two years. The COMPAS risk score for this defendant is [COMPAS SCORE]: [RISK LEVEL].`

The descriptive paragraph in the Advisement treatment provides the following information below the COMPAS score:

> `Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism.`

See Figure A.5-A-7 in the Appendix for an example profile from each treatment group.

As in experiment one, after reviewing a defendant profile participants must evaluate the defendant's risk of recidivism by answering the following two questions:

1. `Please indicate the likelihood that this person will commit another crime by assigning them a recidivism risk score: 1 to 4 (low risk), 5 to 7 (medium risk), 8 to 10 (high risk).`

2. `Do you think this person will commit another crime within two years?`

Additionally, as with experiment one, this survey contains five sections: (1) Consent form; (2) Pre-questionnaire; (3) Instructions; (4) Risk Assessments; (5) Post-questionnaire. All sections are identical to those in experiment one except for the tasks in section (4), which are described above.

4.3 Participants

Also like experiment one, this study employs the Qualtrics platform to deploy the experimental task. It recruits 75 respondents for each treatment through Amazon Mechanical Turk. The task description, worker requirements, and payment structure are the same as those in experiment one.

4.4 Quantitative Measures

This section explicitly defines the measures of accuracy and fairness used in the results section. The chapter employs the notion of overall accuracy discussed in Chapter Two (Background) to evaluate the accuracy of participant predictions. This chapter also denotes the treatment groups as c (for Control), s (for Score), and d (for Disclaimer). For each treatment $t \in \{c, s, d\}$, TP_t, TN_t, FP_t, and

FN_t represent the number of true positive, true negative, false positive, and false negative predictions in that group, respectively. The following accuracy and fairness metrics therefore equal:

- Accuracy rate: $A_t = \frac{TP_t+TN_t}{TP_t+TN_t+FP_t+FN_T}$.
- False positive rate: $FPR_t = \frac{FP}{FP+TN}$
- False negative rate: $FNR_t = \frac{FN}{FN+TP}$
- True positive rate: $TPR_t = \frac{TP}{TP+FN} = 1 - FNR$
- True negative rate: $TNR_t = \frac{TN}{FP+TN} = 1 - FPR$
- Positive predictive value: $PPV_t = \frac{TP}{TP+FP}$

The superscripts b and w denote that a measure only pertains to predictions about black (b) or white (w) defendants. TPR_c^b refers to the true positive rate for black defendants in the Control treatment, and FPR_d^w refers to the false positive rate for white defendants in the Disclaimer treatment, for example.

4.5 Results

75 Mechanical Turk workers participate in each treatment over the course of two weeks in February 2019 for a total of 225 workers. Of this entire set of participants 42.7% list their gender as female, 57.8% list their gender as male, and less than 1% list another option. As with the first experiment, most workers identify as white (79.1%), most identify college as their highest level of education (58.2%), and most specify their age as under 51 (86.2%). When ranking their familiarity with the criminal justice system, most participants report "somewhat familiar" or "familiar," which map to values of 3 and 4 respectively, for a mean result of 3.27

	Average Score	Overall Accuracy	FPR	FNR	PPV
COMPAS Algorithm	4.33	65.0%	30.0%	40.0%	66.7%
Control Treatment	4.99	54.2%	49.8%	36.7%	35.8%
Score Treatment	4.90	51.0%	55.0%	35.1%	34.7%
Disclaimer Treatment	4.79	53.5%	49.2%	40.0%	35.6%

Table 4.1: Performance metrics for recidivism predictions of COMPAS algorithm and Control, Score, and Disclaimer treatment groups.

(compared to mean of 3.11 in the first experiment). Similarly, when ranking their familiarity with machine learning, most participants report "somewhat familiar" or "not very familiar" for a mean result of (2.92) (compared with mean of 2.77 in experiment one). These demographic traits do not significantly differ across participants assigned to different treatments. Table A.2 in the Appendix presents a summary of the participant demographics.

4.5.1 Accuracy

The results suggest that the provision of COMPAS scores does not significantly affect the overall accuracy human predictions of recidivism. In this experiment, $A_c > A_s$: the overall accuracy of predictions among participants who do not see the defendant COMPAS scores (54.2%) exceed that of participants who view the COMPAS scores (51.0%) (See Figure 4.1). A two-sided t-test suggests that this result indicates a possible trend, but it is not statistically significant ($p = 0.0703$).

In this experiment the inclusion of a written advisement about the limitations of the COMPAS algorithm does not significantly affect the accuracy of human predictions of recidivism. Participants in the Disclaimer treatment achieve an average overall accuracy rate of 53.5%, whereas those in the COMPAS condition achieve one of 51.0%, but a two-sided t-test indicates that this difference is not statistically significant ($p = 0.1492$).

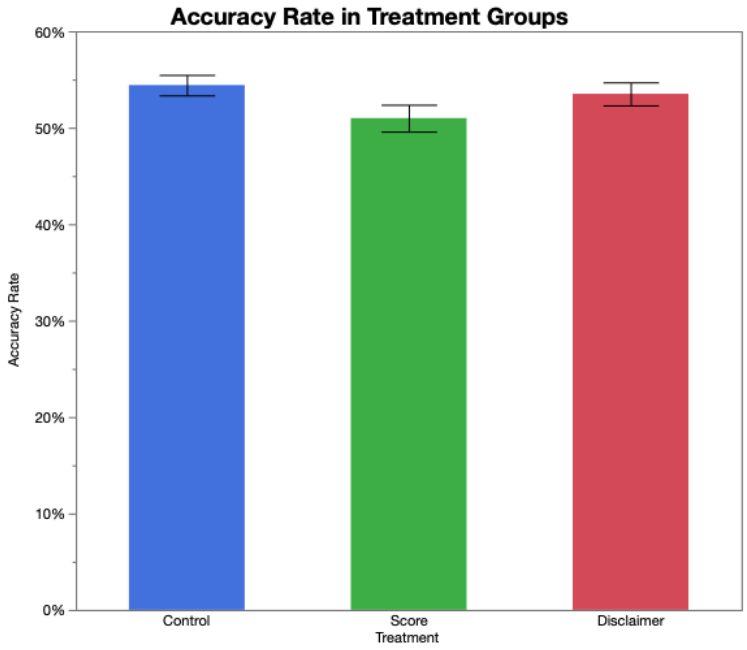


Figure 4.1: Average accuracy of participants in the Control, Score, and Disclaimer treatments. The error bars represent one standard error from the mean.

4.5.2 Confidence

One question in the exit survey of the experiment requires participants to report their confidence in the decisions they made during the task. Respondents answer this question on a 7-point Likert scale. Because Likert scale responses consist of ranked data, this study employs the Wilcoxon/Kruskal-Wallis test to determine if a statistically significant difference exists among the responses in the three conditions. This test indicates that participant confidence does not significantly differ between the Control ($M = 5.84$) and Score ($M = 5.79$) treatments ($p = 0.7696$), between the Control ($M = 5.84$) and Disclaimer ($M = 6.11$) treatments ($p = 0.1256$), or between the Score ($M = 5.79$) and Disclaimer ($M = 6.11$) treatments ($p = 0.0674$).

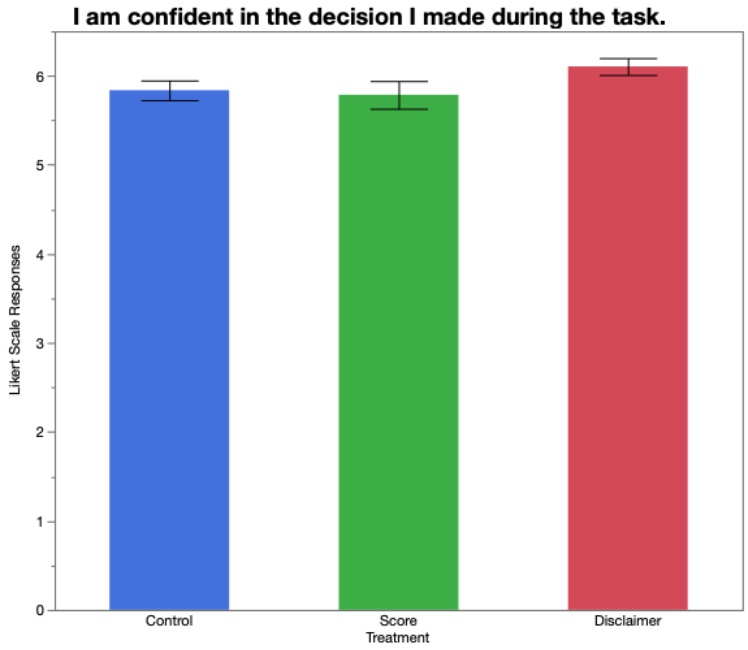


Figure 4.2: Participant reported level of confidence in their decisions during the task. The error bars represent one standard error from the mean.

4.5.3 False Positive and False Negative Rates

The provision of the COMPAS scores significantly increases the false positive rate for black defendants. $FPR_s^b > FPR_c^b$: the false positive rate for black defendants increases from 49.5% in the Control group to 56.5% on the Score group, and a two-sided t-test indicates that this result is statistically significant ($p = 0.0470$). The false positive rate for white defendants does not significantly differ between these treatments ($p = 0.3417$). The Disclaimer does not significantly affect the false positive rate for defendants of either race.

Even though the COMPAS algorithm demonstrates a higher false negative rate for white defendants than the Control group, participant false negative rates for white defendants does not significantly increase when they see the COMPAS score. In fact, a two-sided t-test between the Control and Score treatments indicates that the false negative rate does not significantly differ between the Control

		Average Score	Overall Accuracy	FPR	FNR	TP	TN	PPV
COMPAS Algorithm	Black	4.93	62.9%	41.7%	33.3%	66.7%	58.3%	66.7%
	White	3.08	69.2%	12.5%	60.0%	40.0%	87.5%	66.7%
Control Treatment	Black	4.99	53.9%	49.9%	36.7%	63.3%	50.1%	34.2%
	White	4.99	55.6%	47.5%	38.2%	61.8%	52.5%	39.3%
Compas Treatment	Black	4.96	49.5%	56.5%	35.5%	64.5%	43.5%	32.3%
	White	4.78	54.4%	51.2%	34.3%	65.7%	48.8%	40.4%
Disclaimer Treatment	Black	4.88	52.7%	50.9%	38.5%	61.5%	49.1%	33.2%
	White	4.58	55.5%	47.5%	38.2%	61.8%	44.5%	39.3%

Table 4.2: Performance metrics for recidivism predictions by race of the defendant.

and Score treatments for either white ($p = 0.3527$) or black ($p = 0.7046$) defendants. Again, the Disclaimer does not significantly affect the false negative rate for defendants of either race.

4.5.4 Fairness

The measures of fairness used in the algorithmic fairness literature look beyond the mere magnitude of the false positive and false negative error rates to incorporate more nuanced values. The following section evaluates the predictions in the treatments according to the definitions of fairness as error rate balance, predictive parity, and overall accuracy equity.

4.5.5 Error Rate Balance

According to the balanced error rate notion of fairness, the false positive rates of a predictor should not differ among racial groups, and the false negative rates of a predictor should not differ among racial groups. In the Control group, $FPR_c^b > FPR_c^w$: the false positive rate for black defendants exceeds that of white defendants by 2.4%, but a two-sided t-test indicates that this difference is not statistically significant ($p = 0.1500$). The predictions in the Control thus satisfy

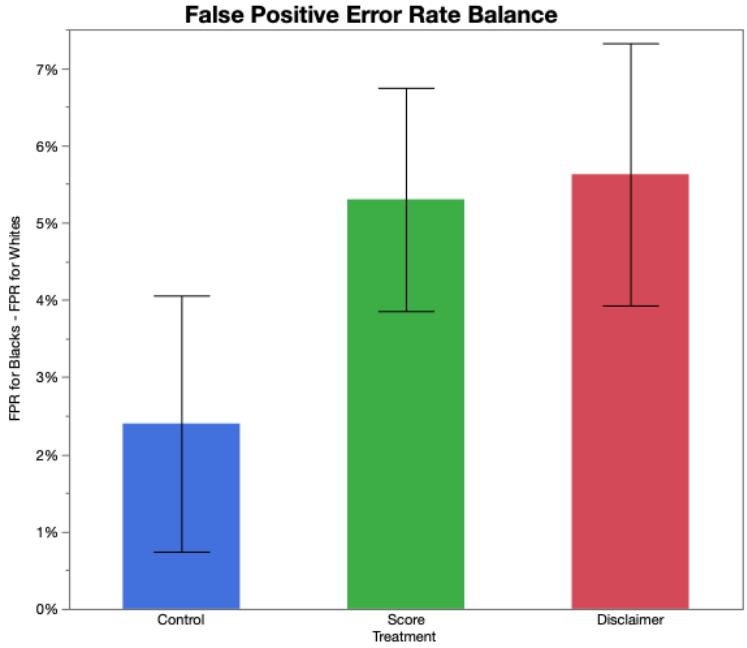


Figure 4.3: The difference between the false positive rate of participant recidivism predictions for black and white defendants in the control group ($FPR_c^b - FPR_c^w$) and the Score ($FPR_s^b - FPR_s^w$) and Disclaimer ($FPR_d^b - FPR_d^w$) treatment groups. The error bars represent one standard error from the mean.

the balanced false positive rates criterion of fairness. When provided with COMPAS scores for the defendants, the false positive rate of participant predictions increases by 5.6% for black defendants but only by 3.7% for white defendants. As such, $FPR_s^b > FPR_s^b$: the false positive rate for black defendants (56.5%) exceeds that for white defendants (51.2%), and this difference is statistically significant ($p = 0.0003$). The predictions in the Score treatment, therefore, do not achieve balanced false positive rates.

The provision of the disclaimer does not significantly affect the difference between the false positive rates for black and white defendants: in this treatment, FPR_d^b significantly decreases to 50.9% ($p = 0.0470$). But FPR_d^w also decreases to 47.5% ($p = 0.1253$). The false positive rate for black thus still significantly exceeds that for white defendants in the disclaimer treatment ($p = 0.0012$). This difference, moreover, does not significantly vary from that in the Score treatment ($p = 0.8858$).

The difference between the false negative rates for black and white defendants is not statistically significant in the Control ($p = 0.5926$), Score ($p = 0.6707$), or Disclaimer ($p = 0.0697$) treatments. Notably, the COMPAS algorithm demonstrates more extreme error rate imbalances than the human predictions. Among this sample of 40 defendants, the algorithm's false positive rate is over three times greater for black defendants (41.7%) than white defendants (12.5%), and the false negative rate is nearly two times greater for white defendants (60.0%) than black defendants (33.3%).

4.5.6 Predictive Parity

The predictions in the Control, Score, and Disclaimer treatments fail to satisfy the criteria for the predictive parity standard of fairness. According to this standard, a predictor's positive prediction value (PPV) must not vary across protected

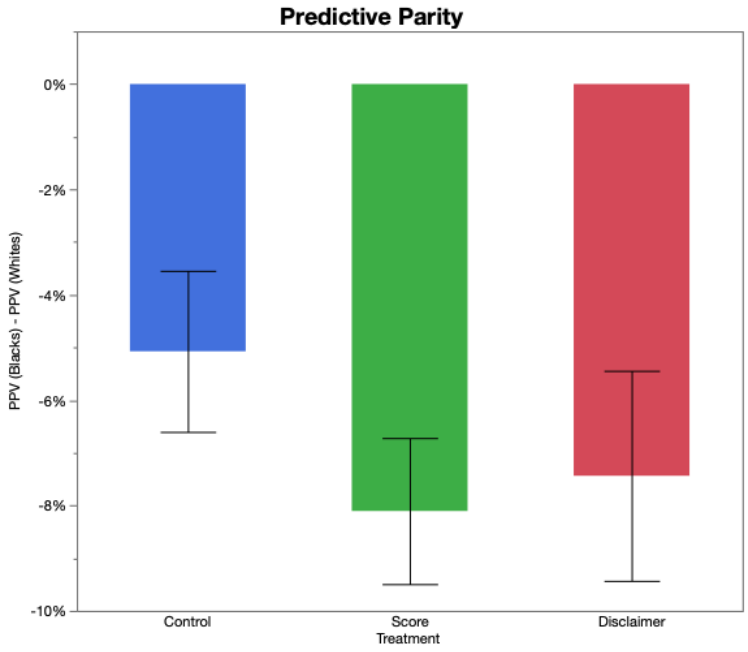


Figure 4.4: The difference between the positive predictive value for predictions about black versus white defendants in the Control ($PPV_c^b - PPV_c^w$), the Score ($PPV_s^b - PPV_s^w$) and the Disclaimer ($PPV_d^b - PPV_d^w$) treatment groups. The error bars represent one standard error from the mean.

groups. In the Control, Score, and Disclaimer treatments, $PPV^w > PPV^b$: the PPV for white defendants exceeds that for black defendants, and two-sided t-tests indicate that these differences are statistically significant with p-values of 0.0011, 0.0000, and 0.0003 respectively. In the Score treatment the PPV for black defendants decreases while the PPV for white defendants increases from that in the Control group, yet a two-sided t-test indicates that the disparity between the positive predictive values ($PPV_s^b - PPV_s^b$) does not significantly differ from that in the Control ($PPV_c^b - PPV_w^b$) group ($p = 0.1998$). Conversely, in the Disclaimer treatment the PPV for black defendants increases while the PPV for white defendants decreases from that in the Score treatment, yet a two-sided t-tests indicates that the disparity between these values does not significantly differ from that in the Score treatment ($p = 0.7755$) or Control group ($p = 0.3156$). In contrast to human predictions, the COMPAS algorithm achieves predictive parity on the sample of defendants used in this experiment with positive predictive values of 66.7% for both white and black defendants.

4.5.7 Overall Accuracy Equity

The predictions in the Control treatment satisfy the criteria for overall accuracy equity, but those in the Score and Disclaimer treatments do not. This notion of fairness mandates that a predictor's accuracy A does not vary across racial groups. The original COMPAS algorithm does not achieve this measure of fairness with an accuracy of 62.9% and 69.2% on this experiment's sample of black and white defendants, respectively. In contrast, two-sided t-tests indicate that A_c^w does not significantly differ from A_c^b: the overall accuracy does not significantly vary between predictions for African-Americans and Caucasians in the Control treatment ($p = 0.2316$). But, in the Score treatment when participants view the defendant COMPAS scores, $A_s^w > A_s^b$: the overall accuracy for white defendants

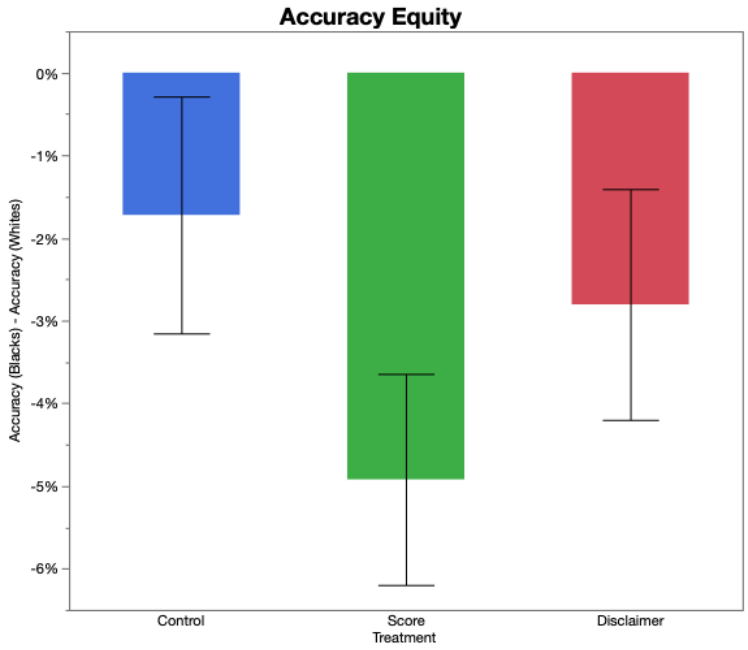


Figure 4.5: The difference between the overall accuracy for predictions about black versus white defendants in the Control group ($A_c^b - A_c^w$) and the Score ($A_s^b - A_s^w$) and Disclaimer ($A_d^b - A_d^w$) treatment groups. The error bars represent one standard error from the mean.

(54.4%) significantly exceeds that for black defendants (49.5%) ($p = 0.0002$). In the Disclaimer treatment when participants view the defendant COMPAS scores with the accompanying written advisement, $A_d^w > A_d^b$: the overall accuracy for white defendants (55.5%) still significantly exceeds that for black defendants (52.7%) ($p = 0.0464$), even though overall accuracy for white defendants decreases while the overall accuracy for black defendants increases.

4.5.8 Other Influences on Decision-Making

The results from this experiment also suggest that participants employ the COMPAS algorithm in their decision-making processes differently depending on their backgrounds. For example, participants who indicate higher levels of familiarity with machine learning report that the COMPAS algorithm influences their de-

cisions more than others ($p < 0.0001$), and they indicated higher confidence in their recidivism predictions, too ($p = 0.0107$). They also indicate higher perceived accuracy ($p < 0.0001$) and fairness ($p = 0.0104$) of the COMPAS algorithm. Notably, a one-way ANOVA test demonstrates that older participants identify significantly lower levels of familiarity with machine learning ($p = 0.0326$). As such, the results also indicate that older participants demonstrate lower perceived influence ($p = 0.0011$), lower perceived accuracy ($p = 0.0089$), and lower perceived fairness ($p = 0.0109$) of the COMPAS algorithm.

4.5.9 Participant Feedback

Upon the conclusion of the task block in the exit survey, 99% of participants responded that they found the instructions for the task clear, and 99% of participants responded that they found the task satisfying. In their feedback, participants indicated they had positive experiences with the study through comments such as: "I thoroughly enjoyed this task," "It was a good length and good payment," and "Very good task."

Participants did not mention the disclaimer when asked how they took the COMPAS scores into account. Rather, participant responses in the Score and Disclaimer treatments demonstrate that they employ the COMPAS scores in non-uniform manners: some ignore them, some rely heavily on them, some use them as a starting point, and others use them as a source of validation. Table 4.3 details excerpts of participant responses.

4.6 Discussion

While experiment one demonstrates that COMPAS scores impact human decision-making by serving as an anchor, this follow-up experiment indicates that provid-

	Compas	Disclaimer
Ignore	"I tried NOT to look at them after awhile, because I felt some were off (lol) but I still took them into account somewhat. I mostly went with my gut and opinions, though."	"I thought it was fairly random, so I didn't invest much faith in it." "Generally I just ignored it and made my own guess."
Rely Heavily	"I kept my scores within 2 points of the compas score."	"I relied on it, it eliminates bias."
Starting Point	"I used that as a baseline."	"It was only a starting point. I paid more attention to the criminal charge, prior charges and guessing on whether the defendant would be convicted." I took a look at the risk score to ballpark it.
Validation	"I used to judge my final answer." "I compared my answer to their answer and adjusted mine slightly based on theirs."	"I made my own guess on what I thought and then checked the compas to find a score I was happy with." "I used the Compas score to verify my decision."
Factor	"I used it in combination with other factors to make my ultimate decision." "I used them in consideration with the other data provided."	"I tended to look at them but used the seriousness of crime or amount of past crimes and age as my main deciding factor." "I took it into consideration along with all the other information given."

Table 4.3: Summary of answers to the free-response question: *How did you incorporate the COMPAS risk scores into your decisions? (Optional).*

ing the algorithm's scores does not significantly affect the accuracy of human predictions of recidivism. The participants who view the COMPAS scores in the defendant profiles do not predict defendant recidivism with significantly different accuracy from those who do not view the risk scores. On the other hand, participants satisfy different fairness criteria from the COMPAS algorithm, even when they consider the algorithm's risk scores in making their predictions. Additionally, as the Wisconsin Supreme Court mandated that a specific written advisement accompany the provision of all COMPAS scores, the experiment also evaluates the impact of including this information alongside the algorithm's risk scores. The findings from this experiment, however, indicate that the advisement does not significantly impact either the accuracy of recidivism predictions or their fairness according to measures of balanced error rates, predictive parity, and accuracy equity. This section readdresses the research hypotheses of the study and contextualizes the results against the backdrop of human-computer interaction and algorithmic fairness literature.

4.6.1 Impact of the Algorithm

When assessing the risk that a defendant will recidivate, the COMPAS algorithm achieves a significantly higher accuracy rate than participants who assess defendant profiles (65.0% v 54.2%). Notably, however, the participants in the control group for this experiment predict recidivism with lower accuracy than those presented with analogous information in Dressel's study [19]. In her experiment participants achieve an overall accuracy rate of 66.5%. Since this study employs Dressel's work as a model, participants view defendant profiles in the same vignette format. The job's task title and description on Amazon Mechanical Turk are identical, too. As such, the difference from Dressel's findings likely result from either (1) the different defendant profiles or (2) the different human participants.

This experiment only employs 40 defendant profiles, but shows the same profiles to each participant. Dressel's study employs 1000 defendant profiles, but divides these profiles into 20 subsets of 50 each that participants then view. Due to the larger number of defendants evaluated, only 20 participants viewed each profile in Dressel's study whereas 75 participants viewed each profile in this study. Perhaps the sample of defendants in this study do not form a representative view of those in the data set, or perhaps the accuracy rate in Dressel's study is unduly influenced by outliers.

Nonetheless, the results from this study suggest that merely providing humans with algorithms that outperform them in terms of accuracy does not necessarily lead to better outcomes. When participants incorporate the algorithm's risk score into their decision-making process, the accuracy rate of their predictions does not significantly change. Given research in complementary computing that shows coupling human and machine intelligence improves their performance [13] [25] [29] [49], this finding seems counterintuitive. Yet prior work in the domain of autonomous vehicles demonstrates that this type of successful collaboration does always occur. For example, a recent study by Axios shows that autonomous vehicles in which humans can take control demonstrate more accidents than those in which humans cannot [38]. Such results suggest that humans may undermine rather than complement machine intelligence in these domains.

Notably, however, in a pilot study of the experiment the opposite effect occurs: participants in the COMPAS treatment predict recidivism with greater accuracy (62.0%) than participants in the Control treatment (52.0%) ($p < 0.0001$) (See Figure 4.6). The random sample of defendants is the only difference between these two studies. In the pilot study, 53.3% of the sample reoffend, 66.7% receive convictions for felonies, and the COMPAS algorithm achieves an accuracy of 79.6% on the set. But in this experiment, 50.0% of the sample reoffend, 65.0%

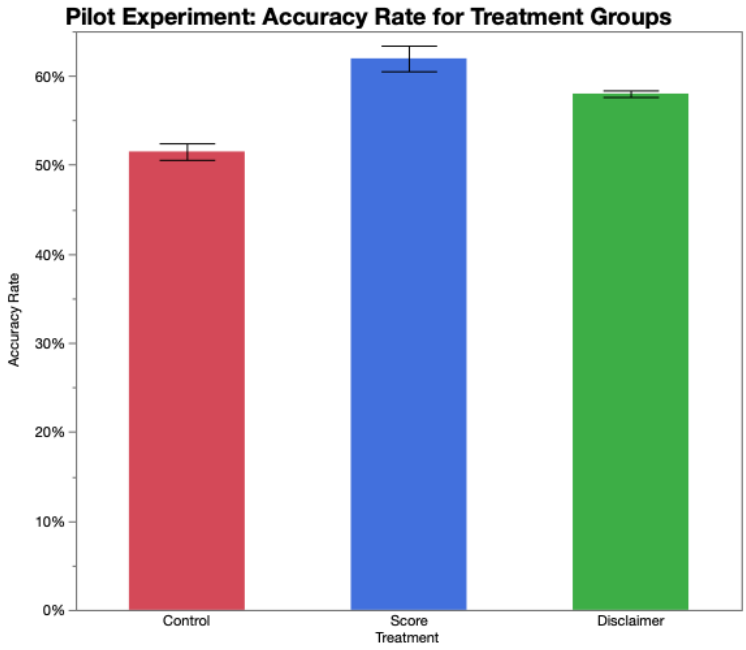


Figure 4.6: Average accuracy of participants in the Control, Score, and Disclaimer treatments in the pilot experiment. The error bars represent one standard error from the mean.

receive convictions for felonies, and the COMPAS algorithm achieves an accuracy of 65.0% on the set. As such, the disparity in results suggests that the accuracy of human predictions of recidivism depend on the profiles of defendants under consideration as well as the accuracy of the COMPAS algorithm on the set.

Yet the results from both experiments highlight the need to target improving human-computer collaboration – not just human or machine judgement in isolation. For example, instead of simply providing humans with the output of an algorithm, people may benefit from tutorials and trainings to familiarize themselves with the tool. During this time, people could also develop the appropriate levels of trust in the algorithm and confidence in their decisions. The results indicate that participants failed to develop the appropriate trust and confidence in their predictions as no significant correlation exists between the confidence level in their predictions and the overall accuracy of their predictions.

The findings from this study also indicate that when humans incorporate an algorithm's outputs as factors into their larger decision-making process, their decisions may satisfy different fairness criteria from those of the algorithm. The prior work in algorithmic fairness focuses on the algorithm in isolation [6] [12] [14] [16] [18] [19] [20], when in practice their outputs serve as mere factors into complex human decisions. For example, the COMPAS algorithm achieves predictive parity on the sample of the defendants in this experiment. But humans who view the COMPAS score when evaluating the defendant profiles do not meet the criteria for predictive parity: the positive predictive value for white defendants significantly exceeds that for blacks. Likewise, for the COMPAS algorithm the false positive rate for black defendants exceeds that for white defendants by 29.2%, yet for participant predictions the false positive rate for black defendants only exceeds that for white defendants by 5.3%. This imbalanced error rate, moreover, does not significantly vary from that among the participants who did not see the

COMPAS scores.

This result could be particularly encouraging for those who fear recidivism prediction algorithms will introduce and/or further embed existing racial biases into the criminal justice system [6]. On the other hand, however, it also serves as a reminder that even when researchers fine tune an algorithm to achieve a specified measure of fairness, human discretion may undermine these efforts. For example, in a study about the effect of algorithmic risk scores on pretrial release decisions, Green finds that participants were 36.4% more likely to deviate positively from the algorithm's risk assessment when evaluating black versus white defendants [26]. Thus just as human discretion may improve fairness according to certain metrics, it also may introduce disparate interactions that arise as humans filter algorithmic predictions.

4.6.2 Impact of the Written Advisement

This study shows that the accuracy of participant predictions of recidivism does not significantly differ when they do and do not see a written advisement about the limitations and appropriate usages of the COMPAS algorithm alongside its output. The Wisconsin Supreme Court mandated the inclusion of this advisement without indicating that they tested its effect on official's decision-making [3]. Psychology research and survey design literature indicate that these types of disclaimers may adversely impact human decision-making in unexpected and unintended manners [27]. But this result suggests that it does not impact the accuracy of decisions in a statistically significant manner.

Notably, however, a pilot study for this experiment, demonstrates a different result. In that study, participants in the Disclaimer treatment underperform those in the COMPAS treatment (58.0% v 62.0%), and a two-sided t-test indicates that this difference is statistically significant ($p = 0.0194$) (see Figure

4.6). The discrepancy between these results, however, could also stem from the different wording of the disclaimer. The advisement in the pilot experiment employed technical language and concrete statistics inspired by ProPublica's study. The section read: "COMPAS scores have been shown to demonstrate accuracy equity across racial groups – the accuracy rate is the same for both white and black defendants. But COMPAS scores have also been shown to demonstrate disparate impact across racial groups – the false positive rate is 45% for black defendants but 25% for white defendants, and the false negative rate is 28% for white defendants but 48% for black defendants." The advisement in this experiment, however, contains the same language the Wisconsin Supreme Court used: "Some studies of COMPAS risk assessment scores have raised questions about whether they disproportionately classify minority offenders as having a higher risk of recidivism." It does not mention or list numbers for false positive rates, false negative rates or positive predictive values. The variance in the results between this study and its pilot could therefore come from their different sample of defendants, participants, or advisements.

While the advisement drafted by the Wisconsin Supreme Court could alleviate disparities between recidivism predictions of black versus white defendants, the results from this experiment indicate that the disclaimer fails to achieve this purpose. As in the Control group and the Score group, in the Disclaimer group predictions of recidivism fail to satisfy the criteria for error rate balance: the false positive rate for black defendants significantly exceeds that for white defendants, and the false negative rate for white defendants significantly exceeds that for black defendants. In practical terms, this disparity could result in more low-risk black defendants incarcerated and more high-risk white defendants set free. The predictions in all three groups also do not meet predictive parity: the probability of a correct positive prediction for white defendants exceeds that for black

defendants. Yet the results in the Control, Score, and Disclaimer groups do not achieve overall accuracy equity: participants more accurately predict recidivism for white defendants than for black defendants.

Taken as a whole, therefore, these metrics indicate that human predictions of recidivism lead to more favorable outcomes for white defendants as opposed to black defendants. But just as the provision of the COMPAS score does not significantly exacerbate the disparate impact between black and white defendants, the provision of the disclaimer does not alleviate it, either. Notably, Chouldechova and Kleinberg demonstrate that the COMPAS algorithm cannot satisfy all of these definitions of fairness – unless, of course, the predictor achieves an overall accuracy rate of 100% [14] [36]. This outcome, however, appears unlikely given the inherent difficulties in predicting human behavior based on limited information, so the designers of recidivism prediction algorithms must make normative decisions and prioritize certain notions of fairness.

4.6.3 Limitations and Future Work

This experiment presents profiles of real defendants from Broward County, Florida, and it randomly samples 40 of them from the data set ProPublica used in their analysis. As such, the measures of accuracy and fairness only pertain to this small subset of individuals. While the overall demographics and criminal histories of the defendants in this sample resemble those in the larger data set, they present a limited view of the combinations and interactions between these characteristics. The various profiles viewed may account for the difference between the observed measures of fairness and accuracy in this experiment and the pilot experiment. Similarly, it may explain why this experiment finds a different baseline accuracy rate for human predictions of recidivism when conducted under analogous circumstances. Future work could conduct this experiment at a larger scale by

receiving evaluations of more defendant profiles – researchers could then explore further relationships between the effect of COMPAS scores and disclaimers on the accuracy and fairness of recidivism predictions among a more robust sample of defendants.

The profiles for black and white defendants in this sample are also not isomorphic: they describe people of different ages, genders, and criminal convictions. While this study suggests that predictions of recidivism result in more favorable assessments of white versus black individuals, other defendant characteristics may also influence the results. Additionally, much like the original data set, this sample contains a greater number of black defendants than white defendants, which could also impact the measures of accuracy and fairness evaluated in the study. As such, future research in a more controlled setting should evaluate the effect of race on predictions of recidivism. For example, researchers could conduct an experiment analogous to this one, but that employs synthetic defendant profiles – namely, they could craft identical vignettes for black and white defendants in which race is the only varying characteristic. With this experimental design modification, they can directly assess the effect of race on predictions of recidivism without confounding factors.

Additionally, as mentioned above, in this study participants view the defendant profiles and COMPAS scores in a vignette format, but in practice criminal justice officials view these scores within the context of the COMPAS software. In the "Person Summary" home screen of the software, they view the defendants name, picture and race, as shown in Figure 4.7 below. This presentation, particularly the defendant's picture, may increase the salience of the defendant's race and thereby lead to greater disparities in treatment by race. In the summary, officials also see the general recidivism risk assessment, but the software reports the score level (i.e. low, medium, high) rather than number. If the officials do not

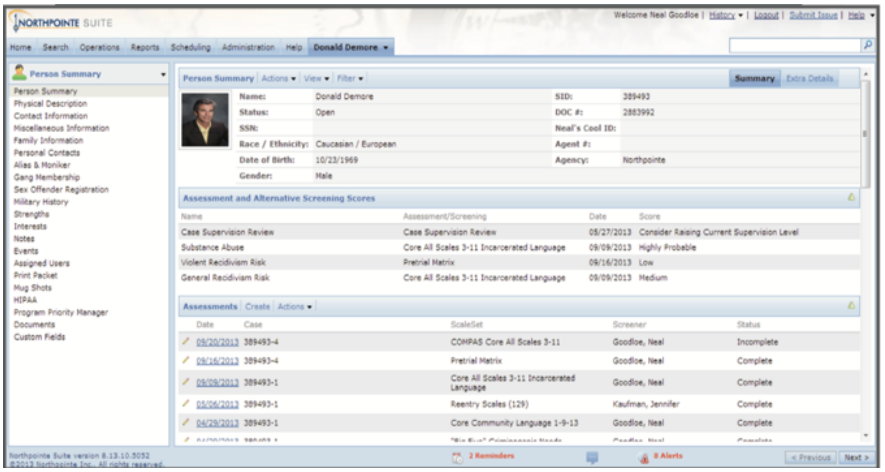


Figure 4.7: "Person Summary" page of the COMPAS software [1]

further evaluate the risk assessment, then they may treat individuals receiving risk scores of 1 and 4 (both low risk) equally. Future work could, therefore, (1) evaluate if the provision of the defendant's image affects racial disparities in recidivism predictions and (2) compare recidivism predictions when presented with risk assessments on a categorical versus numeric scale.

If officials do view a more detailed version of a defendant's risk assessment, they view a bar chart showing the defendant's general recidivism risk, other risk measures for violent recidivism and pretrial release, as well as criminogenic needs such as criminal involvement, relationships, and personality (see Figure 4.8). On the one hand, studies show that these data visualizations can convey information in more accessible, comprehensible, and persuasive manners than traditional text descriptions [47]. On the other hand, people may misunderstand them, particularly when graphs do not show the uncertainty of predictions like the ones in the COMPAS software [48] [58]. Future research should, therefore, investigate how visualizations of algorithmic instruments impact their influence on the accuracy and fairness of recidivism predictions.

Finally, this experiment also faces similar limitations to experiment one as

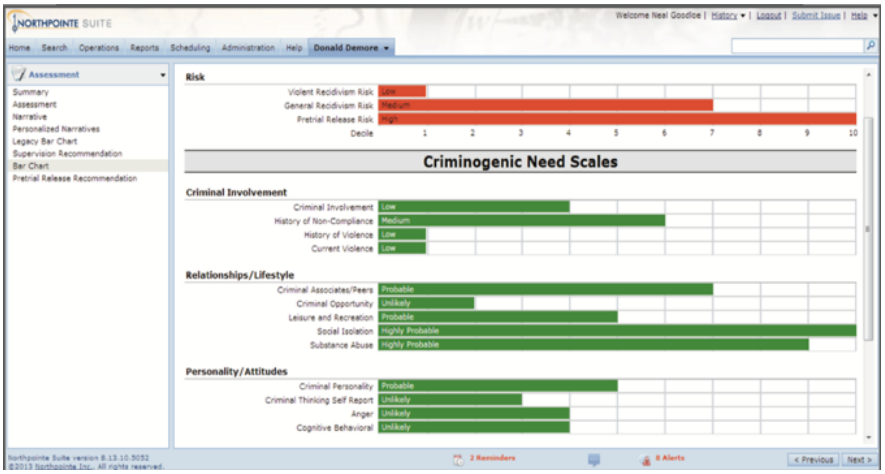


Figure 4.8: Visualization of the defendant risk assessments in the COMPAS software [1]

it too recruits participants through Amazon Mechanical Turk. The respondents lack racial diversity as 79.1% identified as white, and participant race may affect their predictions. The respondents lack training in predictive recidivism, so they may achieve lower accuracy rates than professionals in the criminal justice system. And the respondents may have lacked sufficient incentives to provide their best predictions, so they may not have evaluated each defendant profile to the best of their ability. As such, future work should again strive to employ judges, bail officers, and parole officials as participants to increase the external validity of results.

Nonetheless, the results provide evidence that calls the function of algorithmic risk assessments into question. When humans consider the COMPAS scores, they produce more favorable decisions for white defendants than black defendants according to the notions of balanced error rates and accuracy equity. Additionally, when assessing the risk that a defendant will recidivate, the COMPAS algorithm achieves a significantly higher accuracy rate than the participants who assess defendant profiles (65.0% v 54%). Yet participant accuracy does not increase

when they view the predictions of this more accurate algorithm. Notably, in the pilot study the opposite effect occurs, but this experiment employs a different sample of defendants – the COMPAS algorithm achieves an overall accuracy rate of 80% on the sample. Perhaps a tool with this level of accuracy may, in fact, increase the accuracy of human predictions of recidivism. In the real world, however, no such tool exists; the overall accuracy rate of the COMPAS algorithm is only 65% as in this experimental setting. Thus, this study still raises serious concerns about whether recidivism prediction instruments actually improve outcomes.

Chapter 5

Conclusions

This thesis investigates the role of algorithmic risk assessments in human decision-making using the COMPAS algorithm as a case study. Numerous states employ the COMPAS software to inform decisions about bail, parole, and sentencing. Much of the prior work addressing this algorithm focuses entirely on the technical aspects of the instrument – the tradeoff between its predictive parity and balanced error rates, for example. These studies serve important theoretical purposes, but they do not address the tool's practical function as a decision-making aid rather than decision-maker. Moreover, the few studies that seek to assess the impact of risk assessment software on outcomes in the criminal justice system reach conflicting conclusions, and they employ econometric analyses instead of human subjects [9] [52].

The experiment in Chapter 3, therefore, conducts a controlled study with human subjects to gather quantitative and qualitative measures of the impact of algorithmic risk scores on human predictions of recidivism. The experiment finds that algorithmic risk scores act as anchors that induce a cognitive bias: participants assimilate their predictions to the algorithm's scores. This result echoes studies in psychology that demonstrate home asking-prices, sentencing demands, and dice rolls produce anchoring biases [21] [46]. Like the participants

in those studies, the participants in this experiment do not recognize the degree to which the algorithm's risk scores influence their decisions.

The experiment in Chapter 4 builds upon these results by exploring the follow-up question of how an algorithm's risk scores impact the accuracy and fairness of human decision-making. When predicting the risk that a defendant will recidivate, the COMPAS algorithm achieves a significantly higher accuracy rate than the participants who assess defendant profiles (65.0% v 54.2%). But when participants incorporate the algorithm's risk assessments into their decisions, their accuracy does not improve. In contrast, when participants view the COMPAS scores the fairness of their predictions changes according to measures of balanced error rates and accuracy equity, and they produce more favorable outcomes for white defendants versus black defendants. The Wisconsin Supreme Court designed a written advisement to accompany the presentation of the COMPAS algorithm with the purpose of preventing such adverse impact for minorities. Yet the findings from this experiment indicate that the advisement does not significantly impact either the accuracy of participant recidivism predictions or their fairness according to measures of balanced error rates, predictive parity, and accuracy equity.

Considered in tandem, the results from both experiments underscore the pressing need for further research on the human component of algorithm-aided decision-making in the criminal justice system. Experiment one demonstrates that algorithmic risk assessments influence human predictions about recidivism. Experiment two, however, indicates that the current research on improving accuracy and guaranteeing certain fairness criteria for these instruments does not sufficiently address the nuanced role algorithms play in human assessments of recidivism risk. Even when participants view an algorithm with greater accuracy than them, their performance does not improve. Similarly, when participants in-

corporate an algorithm's outputs as factors into their larger decision-making process, their decisions satisfy different fairness criteria from those of the algorithm. Given human discretion, more accurate and fair algorithms do not necessarily lead to improved decisions or preclude disparate impact on certain groups.

In theory algorithmic risk assessments can alleviate the rate of incarceration and prevent future crimes. They can recognize individuals likely to commit future violent crimes; furthermore, they can help provide people predicted to recidivate with specialized treatment, with the total outcome of reducing crime. The algorithms can also identify low-risk individuals and administer appropriate non-prison punishments, which would help reduce incarceration rates. In practice, however, these risk assessment instruments inform rather than determine outcomes: human officials remain the ultimate arbitrators in bail, parole, and sentencing decisions. Thus, the results from this thesis indicate that if algorithmic risk assessments are to successfully function within the criminal justice system, future research must investigate their practical role as an input to human decision-makers.

Appendix A

Additional Figures and Tables

	Low-Anchor	High-Anchor	Overall
Gender			
Male	31	24	55
Female	18	26	44
Other	1	0	1
Race			
White	44	42	86
Black	0	2	2
Asian	1	4	5
Hispanic	4	1	5
Native American	1	1	2
Other	0	0	0
Education			
PhD	0	0	0
Law School	0	0	0
Master's	8	13	21
College	29	27	56
High School	13	10	23
Did not complete high school	0	0	0
Familiarity with criminal justice system			
Not familiar at all	0	3	3
Slightly familiar	9	9	18
Moderately familiar	28	25	53
Very familiar	10	6	16
Extremely familiar	3	7	10
Familiarity with machine learning			
Not familiar at all	6	8	14
Slightly familiar	17	14	31
Moderately familiar	15	12	27
Very familiar	10	10	20

Table A.1: Participant demographics for experiment one.

	Compas	Control	Disclaimer	Overall
Gender				
Male	41	45	42	128
Female	34	29	32	95
Other	0	0	2	2
Race				
White	55	55	61	171
Black	6	8	6	20
Asian	9	0	3	12
Hispanic	3	2	5	10
Native American	0	2	1	
Other	2	1	0	3
Education				0
PhD	0	0	0	0
Law School	0	1	1	2
Master's	19	13	14	46
College	42	45	43	130
High School	13	15	18	46
Did not complete high school	1	0	0	1
Familiarity with criminal justice system				
Not familiar at all	0	3	3	6
Slightly familiar	15	8	13	36
Moderately familiar	25	34	35	94
Very familiar	28	20	20	68
Extremely familiar	7	9	5	21
Familiarity with machine learning				
Not familiar at all	8	5	10	23
Slightly familiar	15	19	20	54
Moderately familiar	23	31	28	82
Very familiar	22	12	16	50
Extremely Familiar	7	7	2	16

Table A.2: Participant demographics for experiment two.

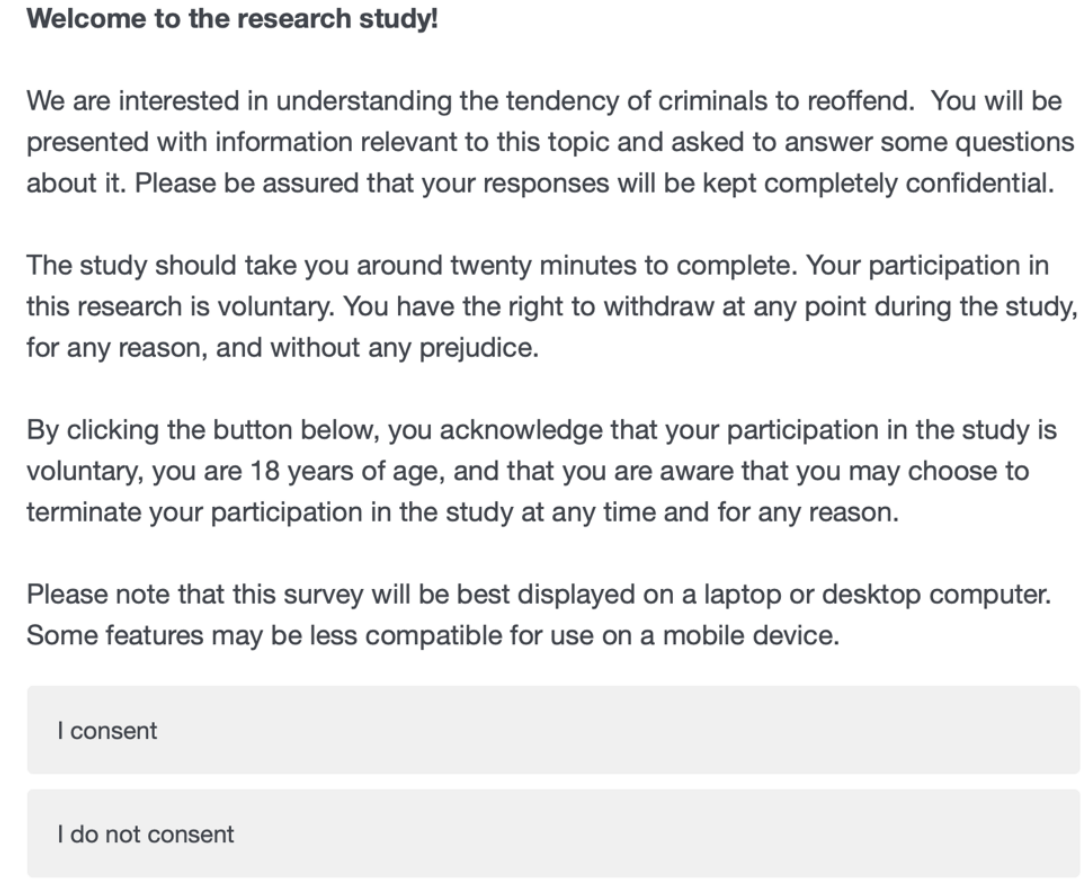


Figure A.1: Survey consent screen.

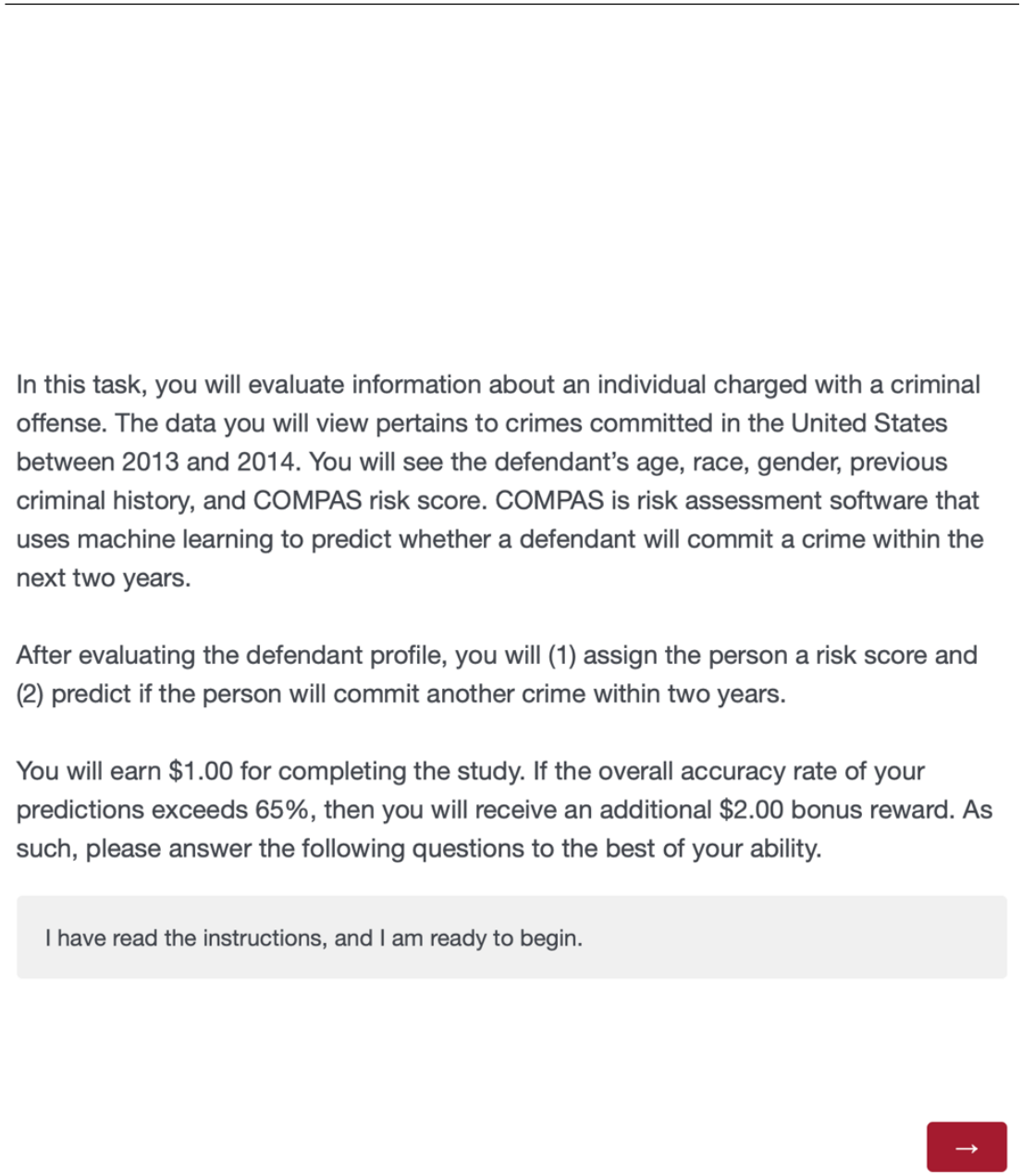


Figure A.2: Survey instructions screen.

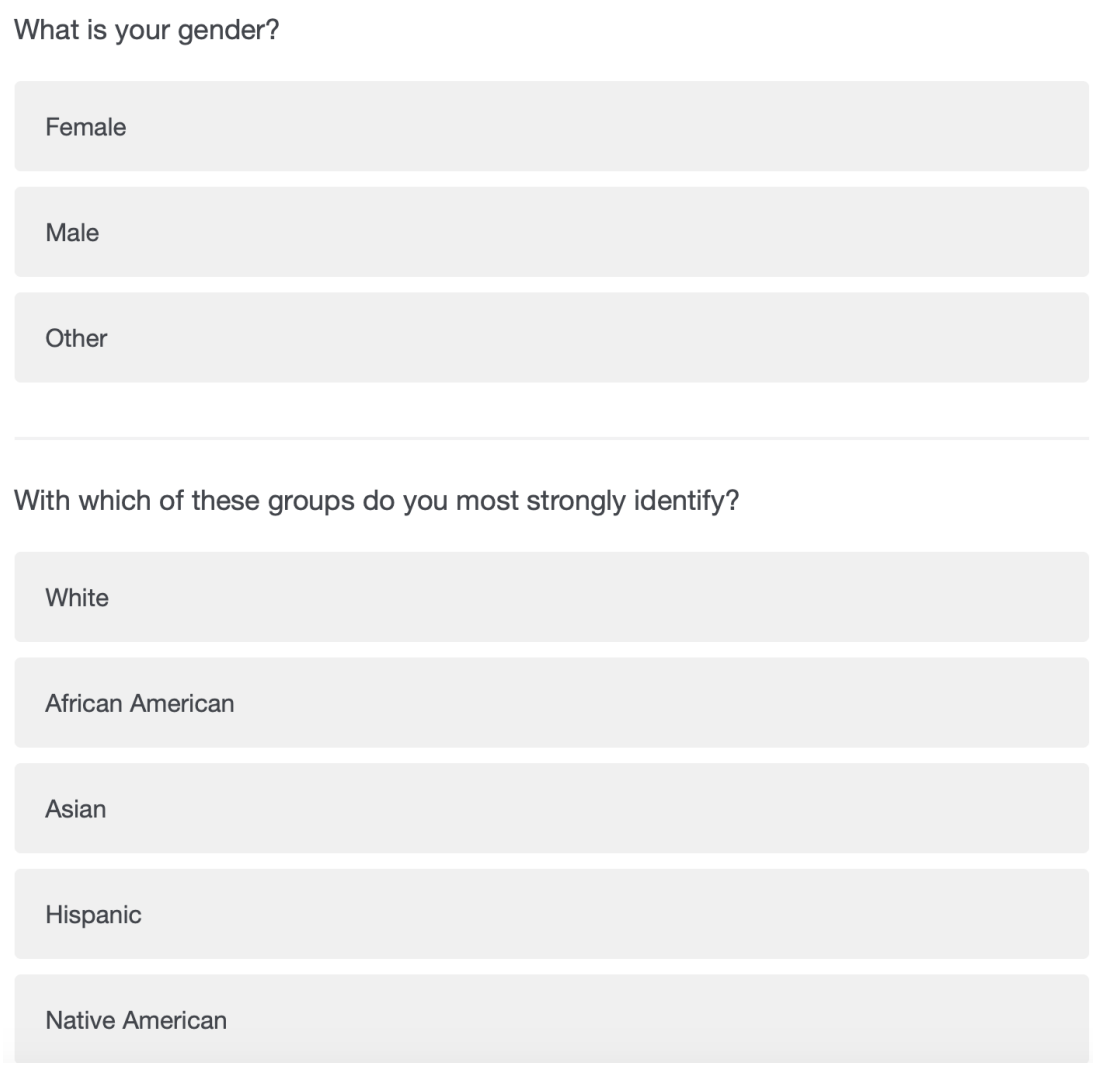


Figure A.3: Survey pre-questionnaire excerpt.

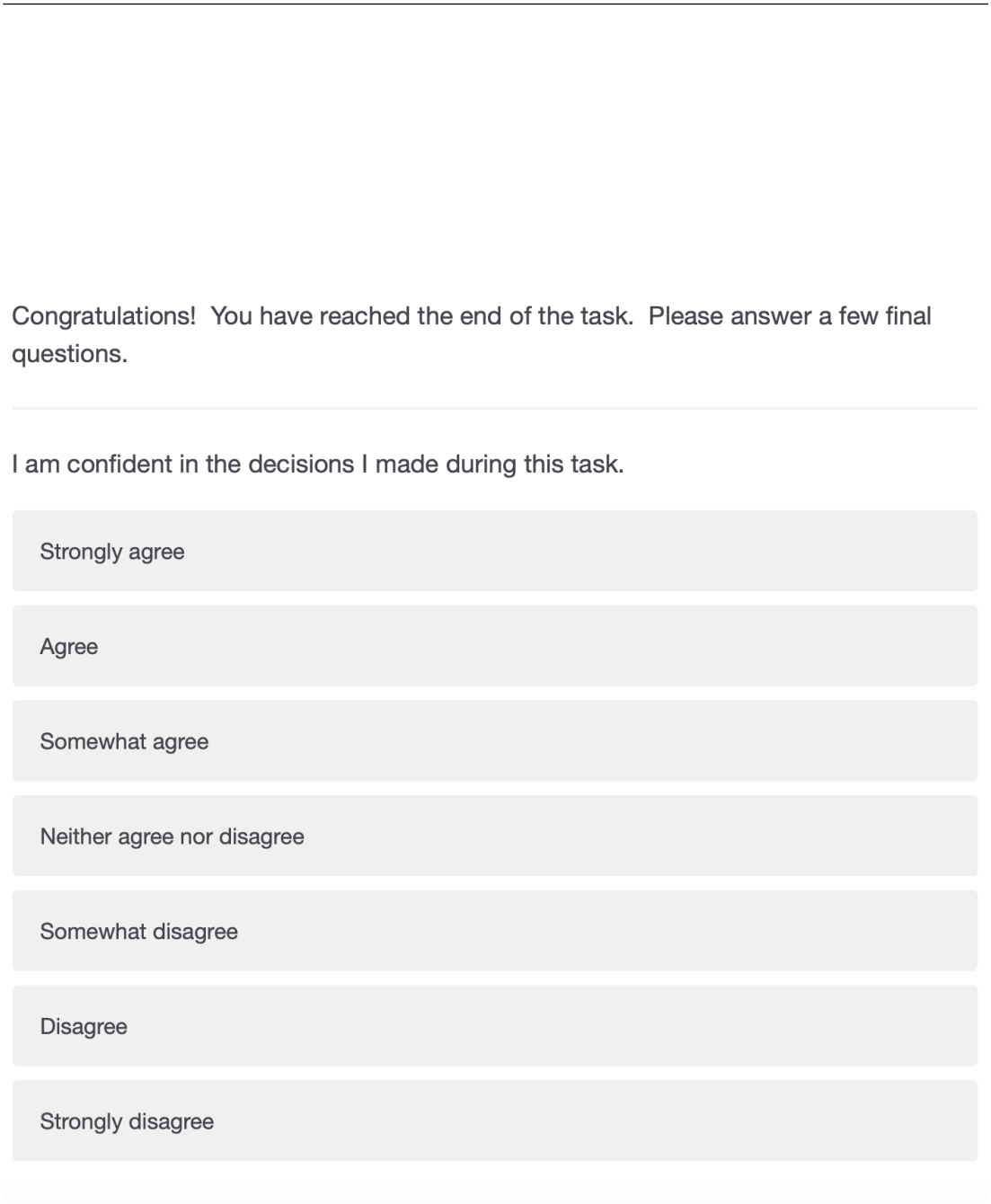


Figure A.4: Survey post-questionnaire excerpt.

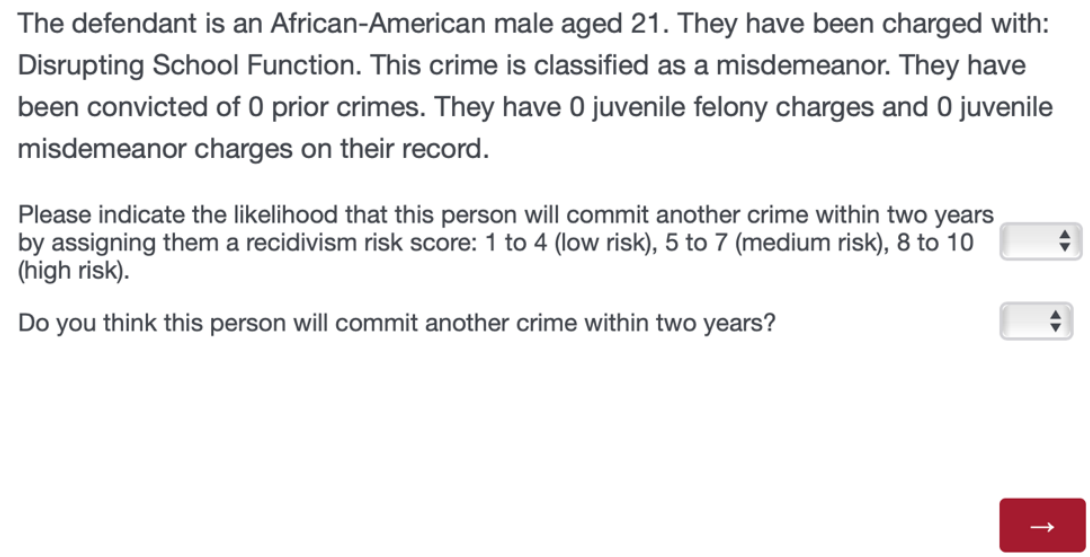


Figure A.5: Example defendant profile in the Control group.

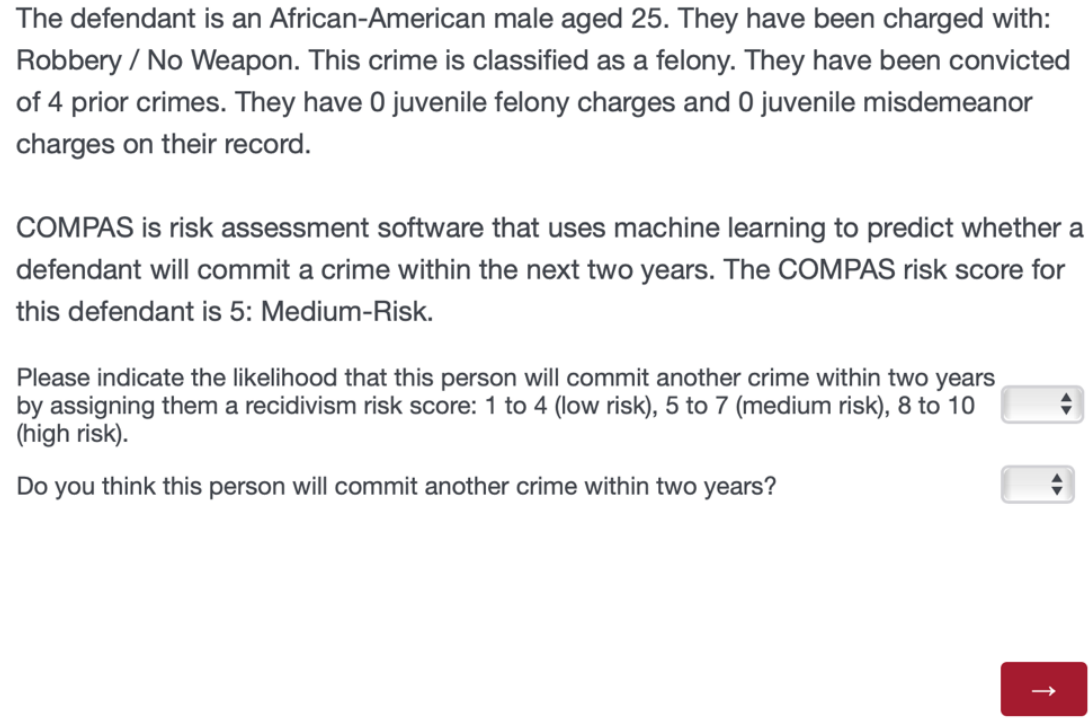


Figure A.6: Example defendant profile in the Score group.

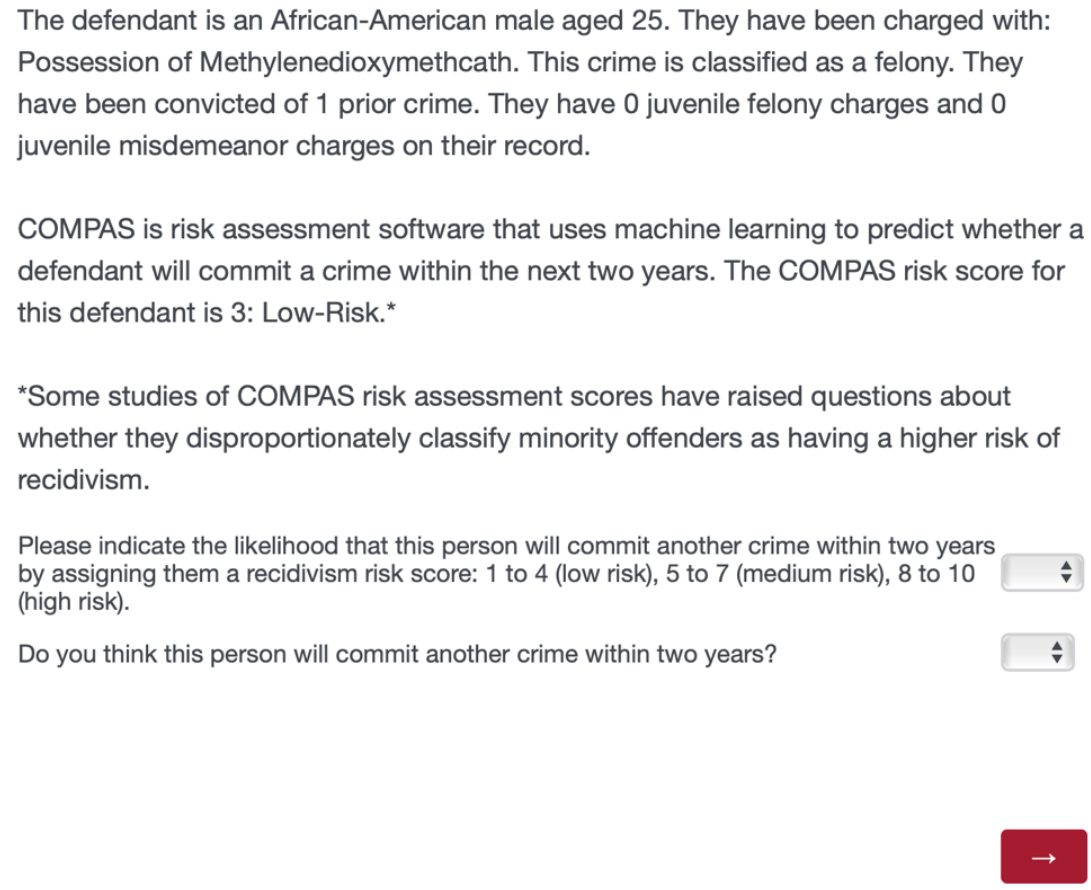


Figure A.7: Example defendant profile in the Disclaimer group.

References

[1] Northpointe suite. vii, 66, 67

[2] United States of America World Prison Brief Data. 1

[3] State v Loomis, 2016. 4, 5, 20, 41, 62

[4] Trends in U.S. Corrections. Technical report, The Sentencing Project, June 2018. 1, 2

[5] D. A. Andrews, James Bonta, and J. Stephen Wormith. The Recent Past and Near Future of Risk and/or Need Assessment. *Crime & Delinquency*, 52(1):7–27, January 2006. 3

[6] Julia Angwin and Jeff Larson. Machine bias. *ProPublica*, May 2016. 3, 4, 12, 13, 14, 22, 23, 42, 61, 62

[7] David Arnold, Will Dobbie, and Crystal S Yang. Racial Bias in Bail Decisions*. *The Quarterly Journal of Economics*, 133(4):1885–1932, November 2018. 2

[8] Jack M. Balkin and Reva B. Siegel. The American Civil Rights Tradition: Anticlassification or Antisubordination. *Issues in Legal Scholarship*, 2(1), January 2003. 13, 14

[9] Richard Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216, June 2017. 9, 69

[10] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research*, page 004912411878253, July 2018. 13, 15, 16, 17, 18

[11] Nicholas A. Bowman and Michael N. Bastedo. Anchoring effects in world university rankings: exploring biases in reputation scores. *Higher Education*, 61(4):431–444, April 2011. 8, 30

[12] Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior*, 36(1):21–40, January 2009. 3, 61

[13] Nicky Case. How To Become A Centaur. *Journal of Design and Science*, January 2018. 11, 21, 59

[14] Alexandra Chouldechova. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data*, 5(2):153–163, June 2017. 5, 13, 15, 16, 17, 18, 21, 24, 61, 64

[15] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*, pages 134–148, January 2018. 19, 21

[16] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the*

23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17, pages 797–806, Halifax, NS, Canada, 2017. ACM Press. 15, 16, 17, 18, 19, 24, 61

[17] Clayton R. Critcher and Thomas Gilovich. Incidental environmental anchors. *Journal of Behavioral Decision Making*, 21(3):241–251, July 2008. 37

[18] William Dieterich, Christina Mendoza, and Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity, July 2016. 12, 13, 26, 61

[19] Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1):eaao5580, January 2018. 5, 10, 21, 24, 28, 39, 58, 61

[20] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, pages 214–226, Cambridge, Massachusetts, 2012. ACM Press. 13, 14, 16, 61

[21] Birte Englich, Thomas Mussweiler, and Fritz Strack. Playing Dice With Criminal Sentences: The Influence of Irrelevant Anchors on Experts Judicial Decision Making. *Personality and Social Psychology Bulletin*, 32(2):188–200, February 2006. 9, 23, 37, 69

[22] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*, pages 259–268, Sydney, NSW, Australia, 2015. ACM Press. 20

[23] Anthony Flores, Christopher Lowenkamp, and Kristin Bechtel. False Positives, False Negatives, and False Analyses: A Rejoinder to Machine Bias: Theres Software Used Across the Country to Predict Future Criminals. And its Biased Against Blacks.. *Federal Probation*, 80(2):38– 46, 2014. 13, 26

[24] Adrian Furnham and Hua Chu Boo. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1):35–42, February 2011. 36

[25] Ian M. Goldstein, Julie Lawrence, and Adam S. Miner. Human-Machine Collaboration in Cancer and Beyond: The Centaur Care Model. *JAMA Oncology*, 3(10):1303, October 2017. 11, 42, 59

[26] Ben Green and Yiling Chen. Disparate Interactions: An Algorithm-in-the-Loop Analysis of Fairness in Risk Assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*, pages 90–99, Atlanta, GA, USA, 2019. ACM Press. 30, 39, 62

[27] Kesten C. Green and J. Scott Armstrong. Evidence on the Effects of Mandatory Disclaimers in Advertising. *Journal of Public Policy & Marketing*, 31(2):293–304, September 2012. 20, 43, 62

[28] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of Opportunity in Supervised Learning. *arXiv:1610.02413 [cs]*, October 2016. arXiv: 1610.02413. 20

[29] Eric Horvitz and Tim Paek. Complementary computing: policies for transferring callers from dialog systems to human receptionists. *User Modeling and User-Adapted Interaction*, 17(1-2):159–182, February 2007. 11, 42, 59

[30] Jonas Jacobson, Jasmine Dobbs-Marsh, Varda Liberman, and Julia A. Minson. Predicting Civil Jury Verdicts: How Attorneys Use (and Misuse) a

Second Opinion: Predicting Civil Jury Verdicts. *Journal of Empirical Legal Studies*, 8:99–119, December 2011. 37

[31] Karen E. Jacowitz and Daniel Kahneman. Measures of Anchoring in Estimation Tasks. *Personality and Social Psychology Bulletin*, 21(11):1161–1166, November 1995. 8, 30

[32] Edward J. Joyce and Gary C. Biddle. Anchoring and Adjustment in Probabilistic Inference in Auditing. *Journal of Accounting Research*, 19(1):120, 1981. 8

[33] Rosalie Jukier. Inside the Judicial Mind: Exploring Judicial Methodology in the Mixed Legal System of Quebec. *European Journal of Comparative Law and Governance*, February 2014. 9

[34] Daniel Kahneman. *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, 1st pbk. ed edition, 2013. OCLC: ocn834531418. 21, 23, 36, 37

[35] Faisal Kamiran, Indr liobait, and Toon Calders. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and Information Systems*, 35(3):613–644, June 2013. 20

[36] Jon Kleinberg. Inherent Trade-Offs in Algorithmic Fairness. *ACM SIGMETRICS Performance Evaluation Review*, 46(1):40–40, January 2019. 5, 19, 64

[37] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. Working Paper 23180, National Bureau of Economic Research, February 2017. 22

[38] Kia Kokalitcheva. People cause most self-driving car accidents in California, August 2018. 59

[39] Min Kyung Lee. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1):2053951718756684, January 2018. 16

[40] Falk Lieder, Thomas L. Griffiths, Quentin J. M. Huys, and Noah D. Goodman. The anchoring bias reflects rational use of cognitive resources. *Psychonomic Bulletin & Review*, 25(1):322–349, February 2018. 8

[41] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, March 2019. 37

[42] Brian S. Lowery, Curtis D. Hardin, and Stacey Sinclair. Social influence effects on automatic racial prejudice. *Journal of Personality and Social Psychology*, 81(5):842–855, 2001. 39

[43] Marlys J. Mason, Debra L. Scammon, and Xiang Fang. The impact of warnings, disclaimers, and product experience on consumers perceptions of dietary supplements. *Journal of Consumer Affairs*, 41(1):74–99, June 2007.

[44] Anne Milgram. Why smart statistics are the key to fighting crime, December 2018. 1, 2, 22

[45] Thomas Mussweiler and Fritz Strack. Numeric Judgments under Uncertainty: The Role of Knowledge in Anchoring. *Journal of Experimental Social Psychology*, 36(5):495–518, September 2000. 9, 37

[46] Gregory B Northcraft and Margaret A Neale. Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing deci-

sions. *Organizational Behavior and Human Decision Processes*, 39(1):84–97, February 1987. 1, 8, 21, 23, 36, 38, 69

[47] Anshul Vikram Pandey, Anjali Manivannan, Oded Nov, Margaret Satterthwaite, and Enrico Bertini. The Persuasive Power of Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2211–2220, December 2014. 66

[48] Anshul Vikram Pandey, Katharina Rall, Margaret L. Satterthwaite, Oded Nov, and Enrico Bertini. How Deceptive are Deceptive Visualizations?: An Empirical Analysis of Common Distortion Techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*, pages 1469–1478, Seoul, Republic of Korea, 2015. ACM Press. 66

[49] P. Jonathon Phillips, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Graldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun-Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White, and Alice J. OToole. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176, June 2018. 11, 42, 59

[50] Aaron D. Shaw, John J. Horton, and Daniel L. Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 conference on Computer supported cooperative work - CSCW '11*, page 275, Hangzhou, China, 2011. ACM Press. 28

[51] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Has-

sabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, December 2018. 11

[52] Megan T. Stevenson. Assessing Risk Assessment in Action. *SSRN Electronic Journal*, 2017. 10, 69

[53] Fritz Strack and Thomas Mussweiler. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of Personality and Social Psychology*, 73(3):437–446, 1997.

[54] Sarah Tan, Julius Adebayo, Kori Inkpen Quinn, and Ece Kamar. Investigating human + machine complementarity for recidivism predictions. *CoRR*, abs/1808.09123, 2018. 12, 39

[55] Jeremy Travis, Bruce Western, and National Research Council (U.S.), editors. *The growth of incarceration in the United States: exploring causes and consequences*. The National Academies Press, Washington, D.C, 2014. 1

[56] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157):1124–1131, September 1974. 8, 30, 37

[57] United States Sentencing Commission. Demographic Differences in Sentencing: An Update to the 2012 *Booker* Report. *Federal Sentencing Reporter*, 30(3):212–229, February 2018. 3

[58] Elisabeth Kersten van Dijk, Wijnand IJsselsteijn, and Joyce Westerink. Deceptive visualizations and user bias: a case for personalization and ambiguity in PI visualizations. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing Adjunct - UbiComp '16*, pages 588–593, Heidelberg, Germany, 2016. ACM Press. 66

[59] Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness - FairWare '18*, pages 1–7, Gothenburg, Sweden, 2018. ACM Press. 13, 14, 15, 16, 17, 18

[60] Brian Wansink, Robert J. Kent, and Stephen J. Hoch. An Anchoring and Adjustment Model of Purchase Quantity Decisions. *Journal of Marketing Research*, 35(1):71, February 1998. 37

[61] Michael Winerip, Michael Schwirtz, and Robert Gebeloff. For Blacks Facing Parole in New York State, Signs of a Broken System. *The New York Times*, December 2016. 2