MIT Case Studies in Social and Ethical Responsibilities of Computing

The Dangers of Risk Prediction in the Criminal Justice System

Julia Dressel¹, Hany Farid²

¹Department of Computer Science, Dartmouth College,

²Department of Electrical Engineering and Computer Sciences, and School of Information,
University of California, Berkeley

Published on: Feb 05, 2021

DOI: 10.21428/2c646de5.f5896f9f

License: [Creative Commons Attribution-NonCommercial 4.0 International License \(CC-BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)

ABSTRACT

Courts across the United States are using computer software to predict whether a person will commit a crime, the results of which are incorporated into bail and sentencing decisions. It is imperative that such tools be accurate and fair, but critics have charged that the software can be racially biased, favoring white defendants over Black defendants. We evaluate the claim that computer software is more accurate and fairer than people tasked with making similar decisions. We also evaluate, and explain, the presence of racial bias in these predictive algorithms.

Julia Dressel

Department of Computer Science, Dartmouth College

Hany Farid

Department of Electrical Engineering and Computer Sciences, and School of Information, University of California, Berkeley

Keywords: algorithmic risk prediction, algorithmic bias, algorithmic fairness, algorithmic transparency, criminal

justice

Author Disclosure(s): This case study summarizes results described in our previously published work: J. Dressel and H. Farid, “[The accuracy, fairness, and limits of predicting recidivism](#),” *Science Advances* 4, no. 1 (January 17, 2018): eaao5580. J. Dressel is now at Recidiviz. The research reported in this case study was completed prior to Dressel’s employment at Recidiviz, and is not an official communication from her employer.

Introduction

We are the frequent subjects of predictive algorithms that determine music recommendations, product advertising, university admission, job placement, and bank loan qualification. In the criminal justice system, predictive algorithms have been used to predict where crimes will most likely occur, who is most likely to commit a violent crime, who is likely to fail to appear at their court hearing, and who is likely to reoffend at some point in the future.¹

Certain types of algorithmic tools known as “risk assessments” have become particularly prevalent in the criminal justice system within the United States. The majority of risk assessments are built to predict recidivism: asking whether someone with a criminal offense will reoffend at some point in the future. These tools rely on an individual’s criminal history, personal background, and demographic information to make these risk predictions.

Various risk assessments are in use across the country to inform decisions at almost every stage in the criminal justice system.² One widely used criminal risk assessment tool, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS, Northpointe), has been used to assess over one million individuals in the criminal justice system since it was developed in 1998.³ The recidivism prediction component of COMPAS—the Recidivism Risk Scale—has been in use since 2000. This software predicts a person’s risk of committing a misdemeanor or felony within two years of assessment from an individual’s demographics and criminal record.⁴

In the past few years, algorithmic risk assessments like COMPAS have become increasingly prevalent in pretrial decision making. In these contexts, an individual who has been arrested and booked in jail is assessed by the algorithmic tool in use by the given jurisdiction. Judges then consider the risk scores calculated by the tool in their decision to either release or detain a criminal defendant before their trial.

In May of 2016, writing for ProPublica, Julia Angwin and colleagues analyzed the efficacy of COMPAS in the pretrial context on over seven thousand individuals arrested in Broward County, Florida, between 2013 and 2014.⁵ The analysis indicated that the predictions were unreliable and racially biased. The authors found that COMPAS’s overall accuracy for white defendants is 67.0%, only slightly higher than its accuracy of 63.8% for Black defendants.⁶ The mistakes made by COMPAS, however, affected Black and white defendants differently: Black defendants who did not recidivate were incorrectly predicted to reoffend at a rate of 44.9%, nearly twice as high as their white counterparts at 23.5%; and white defendants who did recidivate were incorrectly predicted to not reoffend at a rate of 47.7%, nearly twice as high as their Black counterparts at 28.0%. In other words, COMPAS scores appeared to favor white defendants over Black defendants by underpredicting recidivism for white and overpredicting recidivism for Black defendants. Unsurprisingly, this caused an uproar and significant concern that technology was being used to further entrench racism in our criminal justice system.

Since the publication of the ProPublica analysis, there has been significant research and debate regarding the measurement of algorithmic fairness.⁷ Complicating this discussion is the fact that the research community does not necessarily agree on the definition of what makes an algorithm fair. And some studies have revealed that certain definitions of fairness are mathematically incompatible.⁸ To

this date, the debate around mathematical measurement of fairness is both complicated and unresolved.

Algorithmic predictions have become common in the criminal justice system because they maintain a reputation of being objective and unbiased, whereas human decision making is considered inherently more biased and flawed. Northpointe describes COMPAS as “an objective method of estimating the likelihood of reoffending.”⁹ The Public Safety Assessment (PSA), another common pretrial risk assessment tool, advertises itself as a tool to “provide judges with objective, data-driven, consistent information that can inform the decisions they make.”¹⁰ In general, people often assume that algorithms using “big data techniques” are unbiased simply because of the amount of data used to build them.¹¹

After reading the ProPublica analysis in May of 2016, we started thinking about recidivism prediction algorithms and their use in the criminal justice system. To our surprise, we could not find any research proving that recidivism prediction algorithms are superior to human predictions. Due to the serious implications this type of software can have on a person’s life, we felt that we should start by confirming that COMPAS is, in fact, outperforming human predictions. We also felt that it was critical to get beyond the debate of how to measure fairness and understand why COMPAS’s predictive algorithm exhibited such troubling racial bias.

In our study, published in *Science Advances* in January 2018, we began by asking a fundamental question regarding the use of algorithmic risk predictions: are these tools more accurate than the human decision making they aim to replace?¹² The goal of the study was to evaluate the baseline for human performance on recidivism prediction, and assess whether COMPAS was actually outperforming this baseline. We found that people from a popular online crowd-sourcing marketplace—who, it can reasonably be assumed, have little to no expertise in criminal justice—are as accurate and fair as COMPAS at predicting recidivism. This somewhat surprising result then led us to ask: how is it possible that the average person on the internet, being paid \$1 to respond to a survey, is as accurate as commercial software used in the criminal justice system? To answer this, we effectively reverse engineered the COMPAS prediction algorithm and discovered that the software is equivalent to a simple classifier based on only two pieces of data, and it is this simple predictor that leads to the algorithm reproducing historical racial inequities in the criminal justice system.

Comparing Human and Algorithmic Recidivism Prediction

Methodology

Our study is based on a data set of 2013–2014 pretrial defendants from Broward County, Florida.¹³ This data set of 7,214 defendants contains individual demographic information, criminal history, the

COMPAS recidivism risk score, and each defendant's arrest record within a two-year period following the COMPAS scoring, excluding any time spent detained in a jail or a prison. COMPAS scores—ranging from 1 to 10—classify the risk of recidivism as low-risk (1–4), medium-risk (5–7), or high-risk (8–10). For the purpose of binary classification, following the methodology used in the ProPublica analysis and the guidance of the COMPAS practitioner's guide, scores of 5 or above were classified as a prediction of recidivism.¹⁴

Of the 7,214 defendants in the data set, 1,000 were randomly selected for use in our study that evaluated the human performance of recidivism prediction. This subset yields similar overall COMPAS accuracy, false positive rate, and false negative rate as on the complete data set. (A positive prediction is one in which a defendant is predicted to recidivate; a negative prediction is one in which they are predicted to not recidivate.) The COMPAS accuracy for this subset of 1,000 defendants is 65.2%. The average COMPAS accuracy on 10,000 random subsets of size 1,000 each is 65.4% (with a 95% confidence interval of [62.6, 68.1]).

A descriptive paragraph for each of 1,000 defendants was generated:

The defendant is a [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-MISDEMEANOR COUNT] juvenile misdemeanor charges on their record.

Perhaps, most notably, we did not specify the defendant's race in this “no race” condition. In a follow-up “race” condition, the defendant's race was included so that the first line of the above paragraph read, “The defendant is a [RACE] [SEX] aged [AGE].”

There was a total of sixty-three unique criminal charges, including armed robbery, burglary, grand theft, prostitution, robbery, and sexual assault. The crime degree is either “misdemeanor” or “felony.” To ensure that our participants understood the nature of each crime, the above paragraph was followed by a short description of each criminal charge:

[CRIME CHARGE]: [CRIME DESCRIPTION]

After reading the defendant description, participants were then asked to respond either “Yes” or “No” to the question “Do you think this person will commit another crime within two years?” Participants were required to answer each question and could not change their response once it was made. After each answer, participants were given two forms of feedback: whether their response was correct and their average accuracy.

The 1,000 defendants were randomly divided into 20 subsets of 50 each. Each participant was randomly assigned to see one of these 20 subsets. Participants saw the 50 defendants—one at a time—in random order. Participants were only allowed to complete a single subset of 50 defendants.

Participants were recruited through Amazon's Mechanical Turk, an online crowd-sourcing marketplace where people are paid to perform a wide variety of tasks. (Institutional review board [IRB] guidelines were followed for all participants.) Our task was titled "Predicting Crime" with the description "Read a few sentences about an actual person and predict if they will commit a crime in the future." The keywords for the task were "survey, research, criminal justice." Participants were paid one dollar for completing the task and an additional five-dollar bonus if their overall accuracy on the task was greater than 65%. This bonus was intended to provide an incentive for participants to pay close attention to the task. To filter out participants who were not paying close attention, three catch trials were randomly added to the subset of 50 questions. These questions were formatted to look like all other questions but had easily identifiable correct answers.¹⁵ A participant's response was eliminated from our analysis if any of these questions were answered incorrectly.

Responses for the first (no-race) condition were collected from 462 participants, 62 of whom were removed due to an incorrect response on a catch trial. Responses for the second (race) condition were collected from 449 participants, 49 of whom were removed due to an incorrect response on a catch trial. In each condition, this yielded 20 participant responses for each of 20 subsets of 50 questions. Because of the random pairing of participants to a subset of 50 questions, we occasionally oversampled the required number of 20 participants. In these cases, we selected a random 20 participants and discarded any excess responses.

Results

We compare the overall accuracy and bias in human assessment with the algorithmic assessment of COMPAS. Throughout, a *positive* prediction is one in which a defendant is predicted to recidivate while a *negative* prediction is one in which they are predicted to not recidivate. We measure overall accuracy as the rate at which a defendant is correctly predicted to recidivate or not (i.e., the combined true positive and true negative rates). We also report on *false positives* (a defendant is predicted to recidivate but they don't) and *false negatives* (a defendant is predicted to not recidivate but they do). Throughout, we use both paired and unpaired *t*-tests (with 19 degrees of freedom) to analyze the performance of our participants and COMPAS.

The mean and median accuracy in the no-race condition—computed by analyzing the average accuracy of the 400 human predictions—is 62.1% and 64.0%. We compare these results with the performance of COMPAS on this subset of 1,000 defendants. Because groups of 20 participants judged the same subset of 50 defendants, the individual judgments are not independent. However, because each participant

judged only one subset of the defendants, the median accuracies of each subset can reasonably be assumed to be independent. The participant performance, therefore, on the 20 subsets can be directly compared to the COMPAS performance on the same 20 subsets. A one-sided t -test reveals that the average of the 20 median participant accuracies of 62.8% is, just barely, lower than the COMPAS accuracy of 65.2% ($p = 0.045$).

To determine if there is “wisdom in the crowd” (in our case, a small crowd of 20 people per subset), participant responses were pooled within each subset using a majority rules criterion. This crowd-based approach yields a prediction accuracy of 67.0%. A one-sided t -test reveals that COMPAS is not significantly better than the crowd ($p = 0.85$). This demonstrates that the commercial COMPAS prediction algorithm does not outperform small crowds of non-experts at predicting recidivism.

As we noted earlier, there exists significant debate regarding the measurement of algorithmic fairness. For the purpose of this study, we evaluate the human predictions with the same fairness criteria used in the ProPublica analysis for ease of comparability. We acknowledge that this may not be the ideal measure of fairness, and also acknowledge that there is debate in the literature on the appropriate measure of fairness.¹⁶ Regardless, we consider fairness in terms of disparate false positive rates (incorrectly classifying a defendant as high risk when they are not) and false negative rates (incorrectly classifying a defendant as low risk when they are not). We believe that, while perhaps not perfect, this measure of fairness shines a light on real-world consequences of incorrect predictions by quantifying the number of defendants that are improperly incarcerated or released.

We measure the fairness of our participants with respect to a defendant’s race based on the crowd predictions. Our participants’ accuracy on Black defendants is 68.2% compared to 67.6% for white defendants. An unpaired t -test reveals no significant difference across race ($p = .87$). This is similar to that of COMPAS, having a statistically insignificant difference in accuracy of 64.9% for Black defendants and 65.7% for white defendants. By this measure of fairness, our participants and COMPAS are fair to Black and white defendants.

Despite this fairness in overall accuracy, our participants had a significant difference in the false positive and false negative rates for Black and white defendants. Specifically, our participants’ false positive rate for Black defendants is 37.1% compared to 27.2% for white defendants, and our participants’ false negative rate for Black defendants is 29.2% compared to 40.3% for white defendants.

These discrepancies are similar to that of COMPAS, which has a false positive rate of 40.4% for Black defendants and 25.4% for white defendants, and a false negative rate for Black defendants of 30.9% compared to 47.9% for white defendants. See table 1(a) and (c) and figure 1 for a summary of these results. By this measure of fairness, our participants and COMPAS are similarly unfair to Black defendants, despite—bizarrely—the fact that race is not explicitly specified.

Figure 1. Human (no race condition) v. COMPAS algorithmic predictions (see also Table 1).

The results of this study led us to question how human participants produced racially disparate predictions despite not knowing the race of the defendant. We recruited a new set of 400 participants to repeat the same exercise but this time with the defendant's race included. We wondered if including a defendant's race would reduce or exaggerate the effect of any implicit, explicit, or institutional racial bias.

In this race condition, the mean and median accuracy on predicting whether a defendant would recidivate is 62.3% and 64.0%, nearly identical to the condition where race is not specified, see table 1(a) and (b). The crowd-based accuracy is 66.5%, slightly lower than the condition in which race is not specified, but not significantly so. With respect to fairness, participant accuracy is not significantly different for Black defendants, 66.2%, compared to white defendants, 67.6%. The false positive rate for Black defendants is 40.0% compared to 26.2% for white defendants. The false negative rate for Black defendants is 30.1% compared to 42.1% for white defendants. See table 1(b) for a summary of these results.

Table 1*Human v. COMPAS algorithmic predictions from 1,000 defendants*

	(a)	(b)	(c)
	Human (no race)	Human (race)	COMPAS
Accuracy (overall)	67.0%	66.5%	65.2%
Accuracy (Black)	68.2%	66.2%	64.9%
Accuracy (white)	67.6%	67.6%	65.7%
False positive (Black)	37.1%	40.0%	40.4%
False positive (white)	27.2%	26.2%	25.4%
False negative (Black)	29.2%	30.1%	30.9%
False negative (white)	40.3%	42.1%	47.9%
<i>Note: Overall accuracy is specified as percent correct. See also Figure 1.</i>			

Somewhat surprisingly, including race does not have a significant impact on overall accuracy or fairness. Most interestingly, the exclusion of race does not necessarily lead to the elimination of racial disparities in human recidivism prediction.

At this point in our study, we have confidently seen that the COMPAS predictive software is not superior to nonexpert human predictions. However, we are left with two perplexing questions:

1. How is it that nonexperts are as accurate as a widely used commercial software? and
2. How is it that nonexperts appear to be racially biased even when they don't know the race of the defendant?

We next set out to answer these questions.

Replicating COMPAS's Algorithmic Recidivism Prediction

With an overall accuracy of around 65%, COMPAS and nonexpert predictions are not as accurate as we might want, particularly from the point of view of a defendant whose future lies in the balance. Since nonexperts are as accurate as the COMPAS software, we wondered about the sophistication of the

underlying COMPAS predictive algorithm. This algorithm, however, is not publicized, so we built our own predictive algorithm in an attempt to understand and effectively reverse engineer the COMPAS software.

Methodology

Our algorithmic analysis used the same seven features as described in the previous section, extracted from the records in the Broward County data set. Unlike the human assessment that analyzed a subset of these defendants, the following algorithmic assessment is performed over the entire data set.

Logistic regression is a linear classifier that, in a two-class classification (as in our case), computes a separating hyperplane to distinguish between recidivists and nonrecidivists. A nonlinear support vector machine employs a kernel function—in our case, a radial basis kernel—to project the initial seven-dimensional feature space to a higher dimensional space in which a linear hyperplane is used to distinguish between recidivists and nonrecidivists. The use of a kernel function amounts to computing a nonlinear separating surface in the original seven-dimensional feature space, allowing the classifier to capture more complex patterns between recidivists and nonrecidivists than is possible with linear classifiers.

We employed two different classifiers: logistic regression (a simple, general-purpose, linear classifier) and a support vector machine (a more complex, general-purpose, nonlinear classifier).¹⁷ The input to each classifier was seven features from 7,214 defendants: age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior (nonjuvenile) crimes, crime degree, and crime charge (see previous section). Each classifier was trained to predict recidivism from these seven features. Each classifier was trained 1,000 times on a random 80% training and 20% testing split. We report the average testing accuracy.

Results

We found that a simple linear predictor—logistic regression (LR)—provided with the same seven features as our participants (in the no-race condition), yields similar prediction accuracy as COMPAS's predictive algorithm. As compared to COMPAS's overall accuracy of 65.4%, our LR classifier yields an overall testing accuracy of 66.6%. Our predictor also yields similar results to COMPAS in terms of predictive fairness, see table 2(a) and (d).

Despite using only seven features as input, a standard linear predictor yields similar results to COMPAS's software. We can reasonably conclude, therefore, that COMPAS is employing nothing more sophisticated than a linear predictor, or its equivalent.

To test whether performance was limited by the classifier or by the nature of the data, we trained a more powerful nonlinear support vector machine (SVM) on the same data. Somewhat surprisingly, the SVM yields nearly identical results to the linear classifier, see table 2(c). If the relatively low accuracy of the linear classifier was because the data is not linearly separable, then we would have expected the nonlinear SVM to perform better. The failure to do so suggests the data is not separable, linearly or otherwise.

Lastly, we wondered if using an even smaller subset of the seven features would be as accurate as COMPAS. We trained and tested an LR-classifier on all possible subsets of the seven features. In agreement with the research done by Angelino *et al.*, we show that a classifier based on only two features—age and total number of prior convictions—performs as well as COMPAS, see table 2(b).¹⁸ The importance of these two criteria is consistent with the conclusions of two meta-analysis studies that set out to determine, in part, which criteria are most predictive of recidivism.¹⁹

Table 2 <i>Algorithmic predictions from 7,214 defendants</i>				
	(a)	(b)	(c)	(d)
	LR ₇	LR ₂	NL-SVM	COMPAS
Accuracy (overall)	66.6% [64.4, 68.9]	66.8% [64.3, 69.2]	65.2% [63.0, 67.2]	65.4% [64.3, 66.5]
Accuracy (Black)	66.7% [63.6, 69.6]	66.7% [63.5, 69.2]	64.3% [61.1, 67.7]	63.8% [62.2, 65.4]
Accuracy (white)	66.0% [62.6, 69.6]	66.4% [62.6, 70.1]	65.3% [61.4, 69.0]	67.0% [65.1, 68.9]
False positive (Black)	42.9% [37.7, 48.0]	45.6% [39.9, 51.1]	31.6% [26.4, 36.7]	44.8% [42.7, 46.9]
False positive (white)	25.3% [20.1, 30.2]	25.3% [20.6, 30.5]	20.5% [16.1, 25.0]	23.5% [20.7, 26.5]

False negative (Black)	24.2% [20.1, 28.2]	21.6% [17.5, 25.9]	39.6% [34.2, 45.0]	28.0% [25.7, 30.3]
False negative (white)	47.3% [40.8, 54.0]	46.1% [40.0, 52.7]	56.6% [50.3, 63.5]	47.7% [45.2, 50.2]

Note: Predictions are for (a) logistic regression with seven features; (b) logistic regression with two features; (c) a nonlinear support vector machine with seven features; and (d) the commercial COMPAS software with 137 features. The results in columns (a)–(c) correspond to the average testing accuracy over 1,000 random 80/20 training/testing splits. The values in the square brackets correspond to the 95% bootstrapped (a)–(c) and binomial (d) confidence intervals.

In addition to further elucidating the inner workings of these predictive algorithms, the behavior of this two-feature linear classifier helps us understand how the nonexperts were able to match COMPAS's predictive ability. When making predictions about an individual's likelihood of future recidivism, the nonexperts saw the following seven criteria: age, sex, number of juvenile misdemeanors, number of juvenile felonies, number of prior (nonjuvenile) crimes, current crime degree, and current crime charge. If the algorithmic classifier can rely only on a person's age and number of prior crimes to make this prediction, it is plausible that the nonexperts implicitly or explicitly focused on these criteria as well. (Recall that participants were provided with feedback on their correct and incorrect responses, so it is likely that some learning occurred.)

The two-feature classifier effectively learned that if a person is young and has already been convicted multiple times, they are at a higher risk of reoffending, but if a person is older and has not previously been convicted of a crime, then they are at a lower risk of reoffending. This certainly seems like a sensible strategy, if not a terribly accurate one.

The predictive strength of a person's age and number of prior convictions in this context also helps explain the racially disparate predictions seen in both of our human studies and in COMPAS's predictions overall. On a national scale, Black people are more likely to have prior convictions on their record than white people are: for example, Black people in the United States are incarcerated in state prisons at a rate that is 5.1 times that of white Americans.²⁰ Within the data set used in the study, white defendants have an average of 2.59 prior convictions, whereas Black defendants have an average of 4.95 prior convictions. In Florida, the state in which COMPAS was validated for use in Broward County, the incarceration rate of Black people is 3.6 times higher than that of white people.²¹ These racially disparate incarceration rates are not fully explained by different rates of offense by race. Racial disparities against Black people in the United States also exist in policing, arrests, and sentencing.²² The racial bias that appears in both the algorithmic and human predictions is a result of these discrepancies.

While the total number of prior convictions is one of the most predictive variables of recidivism, its predictive power is not very strong. Because COMPAS and the human participants are only moderately accurate (both achieve an accuracy of around 65%), they both make significant, and racially biased, mistakes. Black defendants are more likely to be classified as medium or high risk by COMPAS, because Black defendants are more likely to have prior convictions due to the fact that Black people are more likely to be arrested, charged, and convicted. On the other hand, white defendants are more likely to be classified as low risk by COMPAS, because white defendants are less likely to have prior convictions. Black defendants, therefore, who don't reoffend are predicted to be riskier than white defendants who don't reoffend. Conversely, white defendants who do reoffend are predicted to be less risky than Black defendants who do reoffend. As a result, the false positive rate is higher for Black defendants than white defendants, and the false negative rate for white defendants is higher than for Black defendants. This, in short, is the racial bias that ProPublica first exposed.²³

This same type of disparate outcome appeared in the human predictions as well. Because the human participants saw only a few facts about each defendant, it is safe to assume that the total number of prior convictions was heavily considered in one's predictions. Therefore, the bias of the human predictions was likely also a result of the difference in conviction history, which itself is linked to inequities in our criminal justice system.

The participant and COMPAS's predictions were in agreement for 692 of the 1,000 defendants, indicating that perhaps there could be predictive power in the "combined wisdom" of the risk tool and the human-generated risk scores. However, a classifier that combined the same seven data per defendant along with the COMPAS risk score and the average human-generated risk score performed no better than any of the individual predictions. This suggests that the mistakes made by humans and COMPAS are not independent.

Conclusions

We have shown that a commercial software that is widely used to predict recidivism is no more accurate or fair than the predictions of people with little to no criminal justice expertise who responded to an online survey. We have shown that these predictions are functionally equivalent. When discussing the use of COMPAS in the courtroom to make these life-altering decisions, we should therefore ask whether we would place these same decisions in the equally accurate and biased hands of random people responding to an online survey.

In response to our study, equivant, the makers of COMPAS, responded that our study was both "highly misleading," and "confirmed that COMPAS achieves good predictability."²⁴ Despite this contradictory statement and a promise to analyze our data and results, equivant has not demonstrated any flaws with our study.

Algorithmic predictions—whether in the courts, in university admissions, or employment, financial, and health decisions—can have a profound impact on someone’s life. It is essential, therefore, that the underlying data and algorithms that fuel these predictions are well understood, validated, and transparent to those who are the subject of their use.

In beginning to question the predictive validity of an algorithmic tool, it is essential to also interrogate the ethical implications of the use of the tool. Recidivism prediction tools are used in decisions about a person’s civil liberties. They are, for example, used to answer questions such as “Will this person commit a crime if they are released from jail before their trial? Should this person instead be detained in jail before their trial?” and “How strictly should this person be supervised while they are on parole? What is their risk of recidivism while they are out on parole?” Even if technologists could build a perfect and fair recidivism prediction tool, we should still ask if the use of this tool is just. In each of these contexts, a person is punished (either detained or surveilled) for a crime they have not yet committed. Is punishing a person for something they have not yet done ethical and just?

It is also crucial to discuss the possibility of building any recidivism prediction tool in the United States that is free from racial bias. Recidivism prediction algorithms are necessarily trained on decades of historical criminal justice data, learning the patterns of which kinds of people are incarcerated again and again. The United States suffers from racial discrimination at every stage in the criminal justice system. Machine learning technologies rely on the core assumption that the future will look like the past, and it is imperative that the future of our justice system looks nothing like its racist past. If any criminal risk prediction tool in the United States will inherently reinforce these racially disparate patterns, perhaps they should be avoided altogether.

Discussion Questions

Our study raises important questions as to the appropriateness of using algorithmic prediction in the criminal justice system. At the same time, it also raises a number of questions worthy of thought and debate.

- **Test Yourself:** Nothing will give you a better understanding of the issues at hand than experimenting with the data. From the data that we have made publicly available, it would be relatively straightforward to recreate our experiments and evaluate for yourself how accurate and biased you are at predicting recidivism, and how you compare to algorithmic predictions. All of our experimental data and results may be found at farid.berkeley.edu/downloads/publications/scienceadvances17.
- **Expert vs. Naïve:** We only evaluated the prediction accuracy of naïve participants with no criminal justice experience. Although their accuracy was comparable to algorithmic predictions, it is natural to ask if professionals (attorneys, judges, parole officers, social workers, etc.) would be better or

worse than naïve participants. Shortly after our study was published, we tried to work with one state’s judges to assess their accuracy. We were, however, unable to convince the chief judge to allow us to perform this study (one of the judge’s clerks relayed to us that the chief judge was concerned that the judges would perform no better, or possibly worse, than our naïve participants, and was concerned of the public relations blowback from such a finding). Should judges, prosecutors, police officers, and the like be required to be tested to determine their accuracy and potential bias in parole and sentencing decisions? Do you think this type of training will ameliorate bias and improve fairness in sentencing?

- **Define Fairness:** While we have considered only one measure of fairness, the academic literature has far from agreed upon the most appropriate definition. Researchers at the MIT Technology Review built an interactive “courtroom algorithm game” using the same COMPAS data set used in our study (<https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm>). In this game you can explore the mathematical incompatibilities of various fairness measures. In a criminal justice setting, what definition of fairness is most appropriate? As predictive algorithms are deployed in other settings like university admissions, employment, financial and health decisions, what definitions of fairness should be adopted?
- **Reducing Bias:** If we learn that prior convictions lead to prediction bias, can we simply eliminate this data from our prediction algorithms? Would doing so reduce bias while maintaining accuracy? We again encourage you to experiment with the data available at farid.berkeley.edu/downloads/publications/scienceadvances17 to see if you can make a better and fairer prediction algorithm.
- **Lesser of Two Evils?** In the 2020 election, voters in California had to decide whether to replace California’s existing cash-bail system with algorithmic risk assessment that determines whether a criminal defendant is “too risky” to be released before their trial (<https://voterguide.sos.ca.gov/propositions/25/>). The American Civil Liberties Union (ACLU) of Northern California was “neutral” on this proposition (<https://www.aclunc.org/news/aclu-northern-california-statement-prop-25>). How about you? Californians rejected this measure with 56.4% of the vote. Do you think this was the right decision? The Bail Project has published a “Framework for Reimagining Pretrial Justice” (<https://bailproject.org/after-cash-bail/#pretrial>). Do you agree with their vision? What alternative approaches to pretrial detention do you think would be appropriate?
- **Minority Report:** If you haven’t seen it, we encourage you to watch [Steven Spielberg’s](#) 2002 science fiction thriller, *Minority Report*, loosely based on the 1956 short story “The Minority Report” by Philip K. Dick. Set in the not-too-distant future of 2054, a specialized police unit apprehends suspects based on psychic predictions of the future. Replace psychics with artificial intelligence-based algorithms and ask yourself if we should be arresting people for predictions of their future behavior. Should we adopt such a system if the prediction accuracy is 90%, 95%, 99%, 99.9%? What about 100%?

Bibliography

Angelino, E., N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. "[Learning Certifiably Optimal Rule Lists](#)." In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 35–44. New York: Association for Computing Machinery, 2017.

Angelino, E., N. Larus-Stone, D. Alabi, M. Seltzer, and C. Rudin. "[Learning Certifiably Optimal Rule Lists for Categorical Data](#)." *Journal of Machine Learning Research* 18, no. 234 (2018): 1–78.
<https://arxiv.org/abs/1704.01701>.

Angwin, J., and J. Larson. "[Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say](#)." *ProPublica*, December 30, 2016.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. "[Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased Against Blacks](#)." *ProPublica*, May 23, 2016.

Blomberg, T., W. Bales, K. Mann, R. Meldrum, and J. Nedelec. [Validation of the COMPAS Risk Assessment Classification Instrument](#). Florida State University, Tallahassee, Florida, Technical Report, 2010.

Chouldechova, A. "[Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments](#)." Preprint, submitted February 28, 2017. <https://arxiv.org/abs/1703.00056>.

Corbett-Davies, S., E. Pierson, A. Feller, and S. Goel. "[A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against blacks. It's Actually Not That Clear](#)." *Washington Post*, October 17, 2016.

Dressel, J., and H. Farid. "[The Accuracy, Fairness, and Limits of Predicting Recidivism](#)." *Science Advances* 4, no. 1 (2018): eaao5580.

equivant. "[Official Response to Science Advances](#)." January 18, 2018.

Flores, A. W., K. Bechtel, and C. T. Lowenkamp. "[False Positives, False Negatives, and False Analyses: A Rejoinder to 'Machine Bias: There's Software Used across the Country to Predict Future Criminals. And It's Biased Against Blacks'](#)." *Federal Probation Journal* 80, no. 2 (2016): 38–46.

Gendreau, P., T. Little, and C. Goggin. "[A Meta-analysis of the Predictors of Adult Offender Recidivism: What Works!](#)" *Criminology* 34, no. 4 (1996): 575–608.

Geraghty, K. A., and J. Woodhams. "[The Predictive Validity of Risk Assessment Tools for Female Offenders: A Systematic Review](#)." *Aggression and Violent Behavior* 21 (2015): 25–38.

Hardt, M., E. Price, and N. Srebro. “[Equality of Opportunity in Supervised Learning](#).” Preprint, submitted October 7, 2016. <https://arxiv.org/abs/1610.02413>.

Hastie, R., and T. Kameda. “[The Robust Beauty of Majority Rules in Group Decisions](#).” *Psychological Review* 112 no. 2 (2005): 494–508.

Ho, V. “[Seeking a Better Bail System, SF Turns to Computer Algorithm](#).” *San Francisco Chronicle*, August 1, 2016.

Kleinberg, J., S. Mullainathan, and M. Raghavan. “[Inherent Trade-Offs in the Fair Determination of Risk Scores](#).” Preprint, submitted September 19, 2016, <https://arxiv.org/abs/1609.05807>.

Muñoz, C., M. Smith, and D. J. Patil. [Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights](#). Executive Office of the President, Technical Report. Washington, DC: The White House, May 2016.

Nellis, A. “[The Color of Justice: Racial and Ethnic Disparity in State Prisons](#).” *The Sentencing Project*, June 14, 2016.

Northpointe. [Practitioner’s Guide to COMPAS Core](#). Canton, Ohio, Northpointe Inc., Technical Report, March 2015.

Perry, W. L., B. McInnis, C. C. Price, S. C. Smith, and J. S. Hollywood. [Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations](#). Santa Monica, CA: Rand Corporation, 2013.

Copyright © the Author(s) 2021

Footnotes

1. W. L. Perry et al., [Predictive Policing: The Role of Crime Forecasting in Law Enforcement Operations](#) (Santa Monica, CA: Rand Corporation, 2013). ↵
2. A. Chouldechova, “[Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments](#),” <https://arxiv.org/abs/1703.00056> (submitted February 28, 2017). ↵
3. The software COMPAS is a commercial product of Northpointe, which rebranded itself to “equivant” in January 2017. ↵
4. Northpointe, [Practitioner’s Guide to COMPAS Core](#), Technical Report (March 2015). ↵
5. J. Angwin et al., “[Machine Bias: There’s Software Used across the Country to Predict Future Criminals. And It’s Biased Against Blacks](#),” *ProPublica*, May 23, 2016. ↵

6. With respect to the capitalization of Black, we follow the recommended style guide from the *New York Times*. ↵
7. Chouldechova, “Fair Prediction with Disparate Impact”; S. Corbett-Davies et al., “[A Computer Program Used for Bail and Sentencing Decisions Was Labeled Biased against Blacks. It’s Actually Not That Clear](#),” *Washington Post*, October 17, 2016; J. Kleinberg, S. Mullainathan, and M. Raghavan, “Inherent Trade-Offs in the Fair Determination of Risk Scores,” <https://arxiv.org/abs/1609.05807> arXiv:1609.05807v2 (submitted September 19, 2016); M. Hardt, E. Price, and N. Srebro, “Equality of Opportunity in Supervised Learning,” <https://arxiv.org/abs/1610.02413> (submitted October 7, 2016); A. W. Flores, K. Bechtel, and C. T. Lowenkamp, “[False Positives, False Negatives, and False Analyses: A Rejoinder to ‘Machine Bias: There’s Software Used across the Country to Predict Future Criminals. And It’s Biased against Blacks.’](#)” *Federal Probation Journal* 80, no. 2 (September 2016): 38–46; J. Angwin and J. Larson, “[Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say](#),” *ProPublica*, December 30, 2016. ↵
8. Chouldechova, “Fair Prediction with Disparate Impact”; Kleinberg, Mullainathan, and Raghavan, “Inherent Trade-Offs in the Fair Determination of Risk Scores.” ↵
9. Northpointe, [Practitioner’s Guide to COMPAS Core](#). ↵
10. V. Ho, “[Seeking a Better Bail System, SF Turns to Computer Algorithm](#),” *San Francisco Chronicle*, August 1, 2016. ↵
11. C. Muñoz, M. Smith, and D. J. Patil, [Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights](#), Technical Report (May 2016). ↵
12. J. Dressel and H. Farid, “[The Accuracy, Fairness, and Limits of Predicting Recidivism](#),” *Science Advances* 4, no. 1 (2018): eaao5580. ↵
13. Angwin et al., “[Machine Bias](#).” ↵
14. Angwin et al., “[Machine Bias](#)”; Northpointe, [Practitioner’s Guide to COMPAS Core](#). ↵
15. Catch trials: (1) The state of California was the 31st state to join the Union. California’s nickname is: The Golden State. The state capital is Sacramento. California is bordered by three other states. Los Angeles is California’s most populous city, which is the country’s second largest city after New York City. Does the state of California have a nickname? (2) The first spaceflight that landed humans on the Moon was Apollo 11. These humans were: Neil Armstrong and Buzz Aldrin. Armstrong was the first person to step onto the lunar surface. This landing occurred in 1969. They collected 47.5 pounds of lunar material to bring back to Earth. Did the first spaceflight that landed humans on the

Moon carry Buzz Aldrin? and (3) The Earth is the third planet from the Sun. The shape of Earth is approximately oblate spheroidal. It is the densest planet in the Solar System and the largest of the four terrestrial planets. During one orbit around the Sun, Earth rotates about its axis over 365 times. Earth is home to over 7.4 billion humans. Is Earth the fifth planet from the Sun? [↪](#)

16. Chouldechova, “Fair Prediction with Disparate Impact”; Corbett-Davies et al., “[A Computer Program Used for Bail and Sentencing Decisions](#)”; Kleinberg, Mullainathan, and Raghavan, “Inherent Trade-Offs in the Fair Determination of Risk Scores”; Hardt, Price, and Srebro, “Equality of Opportunity in Supervised Learning”; Flores, Bechtel, and Lowenkamp, “[False Positives, False Negatives, and False Analyses](#)”; Angwin and Larson, “[Bias in Criminal Risk Scores](#).” [↪](#)

17. On logistic regression, see E. Angelino et al., “[Learning Certifiably Optimal Rule Lists for Categorical Data](#),” *Journal of Machine Learning Research* 18 (2018): 1–78, <https://arxiv.org/abs/1704.01701>; on support vector machine, see K. A. Geraghty and J. Woodhams, “[The Predictive Validity of Risk Assessment Tools for Female Offenders: A Systematic Review](#),” *Aggression and Violent Behavior* 21 (2015): 25–38. [↪](#)

18. E. Angelino et al., “[Learning Certifiably Optimal Rule Lists](#),” in *2017 Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Machine Computing, 2017), 35–44. [↪](#)

19. R. Hastie and T. Kameda, “[The Robust Beauty of Majority Rules in Group Decisions](#),” *Psychological Review* 112, no. 2 (2005): 494–508; P. Gendreau, T. Little, and C. Goggin, “[A Meta-analysis of the Predictors of Adult Offender Recidivism: What Works!](#),” *Criminology* 34, no. 4 (November 1996): 575–608. [↪](#)

20. A. Nellis, “[The Color of Justice: Racial and Ethnic Disparity in State Prisons](#),” *The Sentencing Project*, June 14, 2016. [↪](#)

21. T. Blomberg et al., [Validation of the COMPAS Risk Assessment Classification Instrument](#), Technical Report (Broward Sheriff’s Office, September 2010); Nellis, “[The Color of Justice](#).” [↪](#)

22. Nellis, “[The Color of Justice](#).” [↪](#)

23. Angwin et al., “[Machine Bias](#).” [↪](#)

24. equivant, Official response to *Science Advances*, January 18, 2018, <https://www.equivant.com/official-response-to-science-advances/> [↪](#)