

Introduction and Algorithmic Fairness

Responsible Data Science
DS-UA 202 and DS-GA 1017

Compiled by Julia Stoyanovich

This reader contains selected articles on introduction to responsible data science and algorithmic fairness. These articles are part of the required reading for the course. Note that some excerpts end in the middle of a section. Where that is the case, the partial section is not required reading.

Comics: Several comic books are also part of the required reading for this course. Rather than including comics directly into the reader, I include references to the volumes below. These references are to the English-language versions of the comics. Please take a look at the page where all comics are hosted and available for download for other volumes, and for versions in a couple of other languages.

Mirror, Mirror (*Data, Responsibly, vol. 1*) and Who lives, who dies, who decides (*We are AI, vol. 4*) accompany material covered in week 1. I recommend that you read them during the first week of the course. All about that bias (*We are AI, vol. 4*) accompanies material covered through the first module. I recommend that you read it during the second week of the course. Fairness and Friends (*Data, Responsibly, vol. 2*) accompanies material on fairness as equality of opportunity. I recommend that you read it during the fourth week of the course.

Introduction	3
Angwin, Larson, Mattu, and Kirchner (2016) “Machine Bias,” <i>ProPublica</i> .	4
Friedman and Nissenbaum (1996) “Bias in Computer Systems,” <i>ACM Trans. Inf. Syst.</i>	16
Algorithmic Fairness	23
Chouldechova and Roth (2020) “A Snapshot of the Frontiers of Fairness in Machine Learning,” <i>Commun. ACM</i>	24
Kleinberg, Mullainathan, and Raghavan (2016) “Inherent Trade-Offs in the Fair Determination of Risk Scores,” <i>arXiv:1609.05807v2</i>	32
Chouldechova (2017) “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments,” <i>Big Data</i>	39
Obermeyer, Powers, Vogeli, and Mullainathan (2019) “Dissecting Racial Bias in an Algorithm used to Manage the Health of Populations,” <i>Science</i>	46
Arif Khan, Manis, and Stoyanovich (2022) “Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines,” <i>EAAMO</i>	53
Additional Reading	63

Introduction

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica

May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and walked away.

But it was too late — a neighbor who witnessed the heist had already called the police. Borden and her friend were arrested and charged with burglary and petty theft for the items, which were valued at a total of \$80.

Compare their crime with a similar one: The previous summer, 41-year-old Vernon Prater was picked up for shoplifting \$86.35 worth of tools from a nearby Home Depot store.

Prater was the more seasoned criminal. He had already been convicted of armed robbery and attempted armed robbery, for which he served five years in prison, in addition to another armed robbery charge. Borden had a record, too, but it was for misdemeanors committed when she was a juvenile.

Yet something odd happened when Borden and Prater were booked into jail: A computer program spat out a score predicting the likelihood of each committing a future crime. Borden — who is black — was rated a high risk. Prater — who is white — was rated a low risk.

Two years later, we know the computer algorithm got it exactly backward. Borden has not been charged with any new crimes. Prater is serving an eight-year prison term for subsequently breaking into a warehouse and stealing thousands of dollars' worth of electronics.

Subscribe to the Series

Machine Bias: Investigating the algorithms that control our lives.

you@example.com

Subscribe

Read the Documents

- Northpointe document collection
- Sentencing reports that include risk assessments

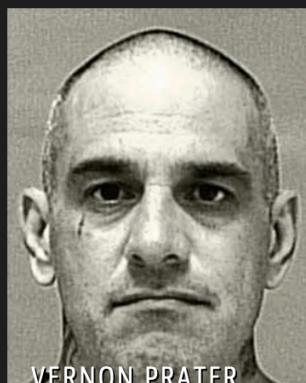
Get the Data

- Read about how we analyzed the risk assessments algorithm
- Download the full data used in our analysis

Scores like this — known as risk assessments — are increasingly common in courtrooms across the nation. They are used to inform decisions about who can be set free at every stage of the criminal justice system, from assigning bond amounts — as is the case in Fort Lauderdale — to even more fundamental decisions about defendants' freedom. In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing.

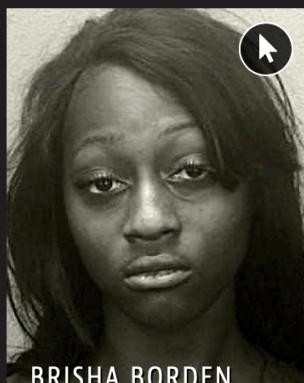
Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation needs. The Justice Department's National Institute of Corrections now encourages the use of such combined assessments at every stage of the criminal justice process. And a landmark sentencing [reform bill](#) currently pending in Congress would mandate the use of such assessments in federal prisons.

Two Petty Theft Arrests



VERNON PRATER

LOW RISK



BRISHA BORDEN

HIGH RISK

In 2014, then U.S. Attorney General Eric Holder warned that the risk scores might be injecting bias into the courts. He called for the U.S. Sentencing Commission to study their use. "Although these measures were crafted with the best of intentions, I am concerned that they inadvertently undermine our efforts to ensure individualized and equal justice," he said, adding, "they may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.

The sentencing commission did not, however, launch a study of risk scores. So ProPublica did, as part of a larger examination of the powerful, largely

hidden effect of algorithms in American life.

We obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over the next two years, the [same benchmark used](#) by the creators of the algorithm.

The score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes actually went on to do so.

When a full range of crimes were taken into account — including misdemeanors such as driving with an expired license — the algorithm was somewhat more accurate than a coin flip. Of those deemed likely to re-offend, 61 percent were arrested for any subsequent crimes within two years.

We also turned up significant racial disparities, just as Holder feared. In forecasting who would re-offend, the algorithm made mistakes with black and white defendants at roughly the same rate but in very different ways.

- The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants.
- White defendants were mislabeled as low risk more often than black defendants.

Could this disparity be explained by defendants' prior crimes or the type of crimes they were arrested for? No. We ran a statistical test that isolated the effect of race from criminal history and recidivism, as well as from defendants' age and gender. Black defendants were still 77 percent more likely to be pegged as at higher risk of committing a future violent crime and 45 percent more likely to be predicted to commit a future crime of any kind. ([Read our analysis.](#))

The algorithm used to create the Florida risk scores is a product of a for-profit company, Northpointe. The company disputes our analysis.

In a letter, it criticized ProPublica's methodology and defended the accuracy of its test: "Northpointe does not agree that the results of your analysis, or the claims being made based upon that analysis, are correct or that they accurately reflect the outcomes from the application of the model."

Northpointe's software is among the most widely used assessment tools in the country. The company does not publicly disclose the calculations used to arrive at defendants' risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity. (On Sunday, Northpointe gave ProPublica the basics of its future-crime formula — which includes factors such as education levels, and whether a defendant has a job. It did not share the specific calculations, which it said are proprietary.)

Northpointe's core product is a set of scores derived from [137 questions](#) that are either answered by defendants or pulled from criminal records. Race is not one of the questions. The survey asks defendants such things as: "Was one of your parents ever sent to jail or prison?" "How many of your friends/acquaintances are taking drugs illegally?" and "How often did you get in fights while at school?" The questionnaire also asks people to agree or disagree with statements such as "A hungry person has a right to steal" and "If people make me angry or lose my temper, I can be dangerous."

ADVERTISEMENT

The appeal of risk scores is obvious: The United States locks up far more people than any other country, a disproportionate number of them black. For more than two centuries, the key decisions in the legal process, from pretrial release to sentencing to parole, have been in the hands of human beings guided by their instincts and personal biases.

If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long. The trick, of course, is to make sure the computer gets it right. If it's wrong in one direction, a dangerous criminal could go free. If it's wrong in another direction, it could result in someone unfairly receiving a harsher sentence or waiting longer for parole than is appropriate.

The first time Paul Zilly heard of his score — and realized how much was riding on it — was during his sentencing hearing on Feb. 15, 2013, in court in Barron County, Wisconsin. Zilly had been convicted of stealing a push lawnmower and some tools. The prosecutor recommended a year in county jail and follow-up supervision that could help Zilly with “staying on the right path.” His lawyer agreed to a plea deal.

But Judge James Babler had seen Zilly’s scores. Northpointe’s software had rated Zilly as a high risk for future violent crime and a medium risk for general recidivism. “When I look at the risk assessment,” Babler said in court, “it is about as bad as it could be.”

Then Babler overturned the plea deal that had been agreed on by the prosecution and defense and imposed two years in state prison and three years of supervision.

CRIMINOLOGISTS HAVE LONG TRIED to predict which criminals are more dangerous before deciding whether they should be released. Race, nationality and skin color were often used in making such predictions until about the 1970s, when it became politically unacceptable, according to a [survey of risk assessment tools](#) by Columbia University law professor Bernard Harcourt.

In the 1980s, as a crime wave engulfed the nation, lawmakers made it much harder for judges and parole boards to exercise discretion in making such decisions. States and the federal government began instituting mandatory sentences and, in some cases, abolished parole, making it less important to evaluate individual offenders.

But as states struggle to pay for swelling prison and jail populations, forecasting criminal risk has made a comeback.

Dozens of risk assessments are being used across the nation — some created by for-profit companies such as Northpointe and others by nonprofit organizations. (One tool being used in states including Kentucky and Arizona, called the Public Safety Assessment, was developed by the Laura and John Arnold Foundation, which also is a funder of ProPublica.)

There have been few independent studies of these criminal risk assessments. In 2013, researchers Sarah Desmarais and Jay Singh [examined 19 different risk methodologies](#) used in the United States and found that “in most cases, validity had only been examined in one or two studies” and that “frequently, those investigations were completed by the same people who developed the instrument.”

Their analysis of the research through 2012 found that the tools “were moderate at best in terms of predictive validity,” Desmarais said in an interview. And she could not find any substantial set of studies conducted in the United States that examined whether risk scores were racially biased. “The data do not exist,” she said.

Two Drug Possession Arrests



DYLAN FUGETT



BERNARD PARKER

LOW RISK

3

HIGH RISK

10

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

Since then, there have been some attempts to explore racial disparities in risk scores. One [2016 study](#) examined the validity of a risk assessment tool, not Northpointe's, used to make probation decisions for about 35,000 federal convicts. The researchers, Jennifer Skeem at University of California, Berkeley, and Christopher T. Lowenkamp from the Administrative Office of the U.S. Courts, found that blacks did get a higher average score but concluded the differences were not attributable to bias.

The increasing use of risk scores is controversial and has garnered media coverage, including articles by the [Associated Press](#), and [the Marshall Project and FiveThirtyEight](#) last year.

Most modern risk tools were originally designed to provide judges with insight into the types of treatment that an individual might need — from drug treatment to mental health counseling.

"What it tells the judge is that if I put you on probation, I'm going to need to give you a lot of services or you're probably going to fail," said Edward Latessa, a University of Cincinnati professor who is the author of a risk assessment tool that is used in Ohio and several other states.

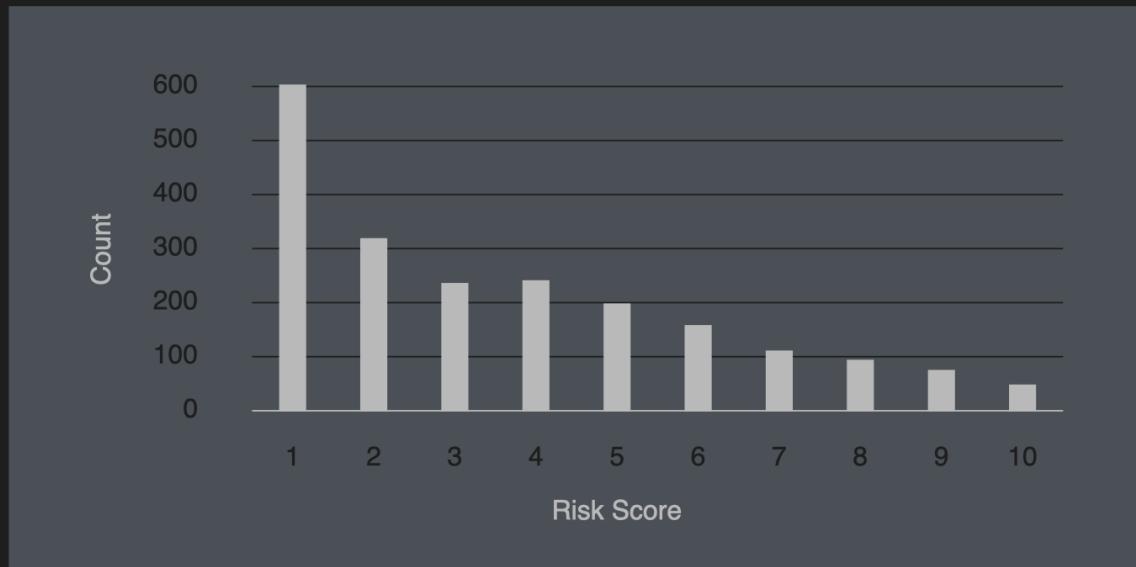
But being judged ineligible for alternative treatment — particularly during a sentencing hearing — can translate into incarceration. Defendants rarely have an opportunity to challenge their assessments. The results are usually shared with the defendant's attorney, but the calculations that transformed the underlying data into a score are rarely revealed.

"Risk assessments should be impermissible unless both parties get to see all the data that go into them," said Christopher Slobogin, director of the criminal justice program at Vanderbilt Law School. "It should be an open, full-court adversarial proceeding."

Black Defendants' Risk Scores



White Defendants' Risk Scores



These charts show that scores for white defendants were skewed toward lower-risk categories. Scores for black defendants were not. (Source: ProPublica analysis of data from Broward County, Fla.)

Proponents of risk scores argue they can be used to reduce the rate of incarceration. In 2002, Virginia became one of the first states to begin using a risk assessment tool in the sentencing of nonviolent felony offenders statewide. In 2014, Virginia judges using the tool sent nearly half of those defendants to alternatives to prison, according to a state sentencing commission report. Since 2005, the state's prison population growth has slowed to 5 percent from a rate of 31 percent the previous decade.

In some jurisdictions, such as Napa County, California, the probation department uses risk assessments to suggest to the judge an appropriate probation or treatment plan for individuals being sentenced. Napa County Superior Court Judge Mark Boessenecker said he finds the recommendations helpful. "We have a dearth of good treatment programs, so filling a slot in a program with someone who doesn't need it is foolish," he said.

However, Boessenecker, who trains other judges around the state in evidence-based sentencing, cautions his colleagues that the score doesn't necessarily reveal whether a person is dangerous or if they should go to prison.

"A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job," Boessenecker said. "Meanwhile, a drunk guy will look high risk because he's homeless. These risk factors don't tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be."

Sometimes, the scores make little sense even to defendants.

James Rivelli, a 54-year old Hollywood, Florida, man, was arrested two years ago for shoplifting seven boxes of Crest Whitestrips from a CVS drugstore. Despite a criminal record that included aggravated assault, multiple thefts and felony drug trafficking, the Northpointe algorithm classified him as being at a low risk of reoffending.

"I am surprised it is so low," Rivelli said when told by a reporter he had been rated a 3 out of a possible 10. "I spent five years in state prison in Massachusetts. But I guess they don't count that here in Broward County."

In fact, criminal records from across the nation are supposed to be included in risk assessments.

Less than a year later, he was charged with two felony counts for shoplifting about \$1,000 worth of tools from Home Depot. He said his crimes were fueled by drug addiction and that he is now sober.



"I'm surprised [my risk score] is so low. I spent five years in state prison in Massachusetts." (Josh Ritchie for ProPublica)

NORTHPOINTE WAS FOUNDED in 1989 by Tim Brennan, then a professor of statistics at the University of Colorado, and Dave Wells, who was running a corrections program in Traverse City, Michigan.

Wells had built a prisoner classification system for his jail. "It was a beautiful piece of work," Brennan said in an interview conducted before ProPublica had completed its analysis. Brennan and Wells shared a love for what Brennan called "quantitative taxonomy" — the measurement of personality traits such as intelligence, extroversion and introversion. The two decided to build a risk assessment score for the corrections industry.

Brennan wanted to improve on a leading risk assessment score, the LSI, or Level of Service Inventory, which had been developed in Canada. "I found a fair amount of weakness in the LSI," Brennan said. He wanted a tool that addressed the major theories about the causes of crime.

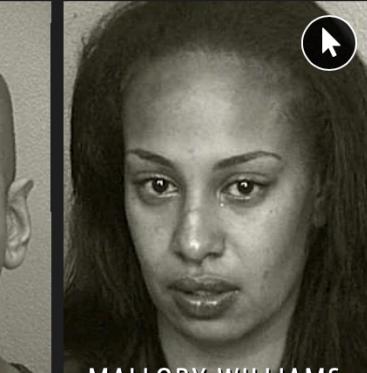
Brennan and Wells named their product the Correctional Offender Management Profiling for Alternative Sanctions, or COMPAS. It assesses not just risk but also nearly two dozen so-called "criminogenic needs" that relate to the major theories of criminality, including "criminal personality," "social isolation," "substance abuse" and "residence/stability." Defendants are ranked low, medium or high risk in each category.

Two DUI Arrests



GREGORY LUGO

LOW RISK



MALLORY WILLIAMS

1

MEDIUM RISK

6

Lugo crashed his Lincoln Navigator into a Toyota Camry while drunk. He was rated as a low risk of reoffending despite the fact that it was at least his fourth DUI.

As often happens with risk assessment tools, many jurisdictions have adopted Northpointe's software before rigorously testing whether it works. New York State, for instance, started using the tool to assess people on probation in a pilot project in 2001 and rolled it out to the rest of the state's probation departments — except New York City — by 2010. The state didn't publish a comprehensive [statistical evaluation](#) of the tool until 2012. The study of more than 16,000 probationers found the tool was 71 percent accurate, but it did not evaluate racial differences.

A spokeswoman for the New York state division of criminal justice services said the study did not

examine race because it only sought to test whether the tool had been properly calibrated to fit New York's probation population. She also said judges in nearly all New York counties are given defendants' Northpointe assessments during sentencing.

In 2009, Brennan and two colleagues [published a validation study](#) that found that Northpointe's risk of recidivism score had an accuracy rate of 68 percent in a sample of 2,328 people. Their study also found that the score was slightly less predictive for black men than white men — 67 percent versus 69 percent. It did not examine racial disparities beyond that, including whether some groups were more likely to be wrongly labeled higher risk.

Brennan said it is difficult to construct a score that doesn't include items that can be correlated with race — such as poverty, joblessness and social marginalization. "If those are omitted from your risk assessment, accuracy goes down," he said.

In 2011, Brennan and Wells sold Northpointe to Toronto-based [conglomerate Constellation Software](#) for an undisclosed sum.

Wisconsin has been among the most eager and expansive users of Northpointe's risk assessment tool in sentencing decisions. In 2012, the Wisconsin Department of Corrections launched the use of the software throughout the state. It is used at each step in the prison system, from sentencing to parole.

In a 2012 presentation, corrections official Jared Hoy described the system as a "[giant correctional pinball machine](#)" in which correctional officers could use the scores at every "decision point."

Wisconsin has not yet completed a statistical validation study of the tool and has not said when one might be released. State corrections officials declined repeated requests to comment for this article.

Some Wisconsin counties use other risk assessment tools at arrest to determine if a defendant is too risky for pretrial release. Once a defendant is convicted of a felony anywhere in the state, the Department of Corrections attaches Northpointe's assessment to the confidential presentence report given to judges, according to Hoy's presentation.

In theory, judges are not supposed to give longer sentences to defendants with higher risk scores. Rather, they are supposed to use the tests primarily to determine which defendants are eligible for probation or treatment programs.

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Overall, Northpointe's assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. (Source: ProPublica analysis of data from Broward County, Fla.)

But judges have cited scores in their sentencing decisions. In August 2013, Judge Scott Horne in La Crosse County, Wisconsin, declared that defendant Eric Loomis had been "identified, through the COMPAS assessment, as an individual who is at high risk to the community." The judge then imposed a sentence of eight years and six months in prison.

Loomis, who was charged with driving a stolen vehicle and fleeing from police, is challenging the use of the score at sentencing as a violation of his due process rights. The state has defended Horne's use of the score with the argument that judges can consider the score in addition to other factors. It has also stopped including scores in presentencing reports until the state Supreme Court decides the case.

"The risk score alone should not determine the sentence of an offender," Wisconsin Assistant Attorney General Christine Remington said last month during state Supreme Court arguments in the Loomis case. "We don't want courts to say, this person in front of me is a 10 on COMPAS as far as risk, and therefore I'm going to give him the maximum sentence."

That is almost exactly what happened to Zilly, the 48-year-old construction worker sent to prison for stealing a push lawnmower and some tools he intended to sell for parts. Zilly has long struggled with a meth habit. In 2012, he had been working toward recovery with the help of a Christian pastor when he relapsed and committed the thefts.

After Zilly was scored as a high risk for violent recidivism and sent to prison, a public defender appealed the sentence and called the score's creator, Brennan, as a witness.

Brennan testified that he didn't design his software to be used in sentencing. "I wanted to stay away from the courts," Brennan said, explaining that his focus was on reducing crime rather than punishment. "But as time went on I started realizing that so many decisions are made, you know, in the courts. So I gradually softened on whether this could be used in the courts or not."



"Not that I'm innocent, but I just believe people do change." (Stephen Muren for ProPublica)

Still, Brennan testified, "I don't like the idea myself of COMPAS being the sole evidence that a decision would be based upon."

After Brennan's testimony, Judge Babler reduced Zilly's sentence, from two years in prison to 18 months. "Had I not had the COMPAS, I believe it would likely be that I would have given one year, six months," the judge said at an appeals hearing on Nov. 14, 2013.

Zilly said the score didn't take into account all the changes he was making in his life — his conversion to Christianity, his struggle to quit using drugs and his efforts to be more available for his son. "Not that I'm innocent, but I just believe people do change."

FLORIDA'S BROWARD COUNTY, where Brisha Borden stole the Huffy bike and was scored as high risk, does not use risk assessments in sentencing. "We don't think the [risk assessment] factors have any bearing on a sentence," said David Scharf, executive director of community programs for the Broward County Sheriff's Office in Fort Lauderdale.

Broward County has, however, adopted the score in pretrial hearings, in the hope of addressing jail overcrowding. A court-appointed monitor has overseen Broward County's jails since 1994 as a result of the settlement of a lawsuit brought by inmates in the 1970s. Even now, years later, the Broward County jail system is often more than 85 percent full, Scharf said.

In 2008, the sheriff's office decided that instead of building another jail, it would begin using Northpointe's risk scores to help identify which defendants were low risk enough to be released on bail pending trial. Since then, nearly everyone arrested in Broward has been scored soon after being booked. (People charged with murder and other capital crimes are not scored because they are not eligible for pretrial release.)

The scores are provided to the judges who decide which defendants can be released from jail. "My feeling is that if they don't need them to be in jail, let's get them out of there," Scharf said.

Scharf said the county chose Northpointe's software over other tools because it was easy to use and produced "simple yet effective charts and graphs for judicial review." He said the system costs about \$22,000 a year.

In 2010, researchers at Florida State University examined the use of Northpointe's system in Broward County over a 12-month period and concluded that its predictive accuracy was "equivalent" in assessing defendants of different races. Like others, they did not examine whether different races were classified differently as low or high risk.

Scharf said the county would review ProPublica's findings. "We'll really look at them up close," he said.

Broward County Judge John Hurley, who oversees most of the pretrial release hearings, said the scores were helpful when he was a new judge, but now that he has experience he prefers to rely on his own judgment. "I haven't relied on COMPAS in a couple years," he said.

Hurley said he relies on factors including a person's prior criminal record, the type of crime committed, ties to the community, and their history of failing to appear at court proceedings.

ProPublica's analysis reveals that higher Northpointe scores are slightly correlated with longer pretrial incarceration in Broward County. But there are many reasons that could be true other than judges being swayed by the scores — people with higher risk scores may also be poorer and have difficulty paying bond, for example.

Most crimes are presented to the judge with a recommended bond amount, but he or she can adjust the amount. Hurley said he often releases first-time or low-level offenders without any bond at all.

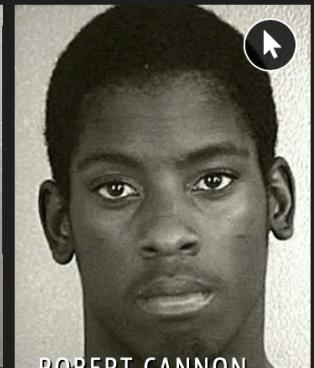
However, in the case of Borden and her friend Sade Jones, the teenage girls who stole a kid's bike and scooter, Hurley raised the bond amount for each girl from the recommended \$0 to \$1,000 each.

Two Shoplifting Arrests



JAMES RIVELLI

LOW RISK



ROBERT CANNON

MEDIUM RISK **6**

After Rivelli stole from a CVS and was caught with heroin in his car, he was rated a low risk. He later shoplifted \$1,000 worth of tools from a Home Depot.

Hurley said he has no recollection of the case and cannot recall if the scores influenced his decision.



Sade Jones, who had never been arrested before, was rated a medium risk. (Josh Ritchie for ProPublica)

The girls spent two nights in jail before being released on bond.

“We literally sat there and cried” the whole time they were in jail, Jones recalled. The girls were kept in the same cell. Otherwise, Jones said, “I would have gone crazy.” Borden declined repeated requests to comment for this article.

Jones, who had never been arrested before, was rated a medium risk. She completed probation and got the felony burglary charge reduced to misdemeanor trespassing, but she has still struggled to find work.

“I went to McDonald’s and a dollar store, and they all said no because of my background,” she said. “It’s all kind of difficult and unnecessary.”



Julia Angwin is a senior reporter at ProPublica. From 2000 to 2013, she was a reporter at The Wall Street Journal, where she led a privacy investigative team that was a finalist for a Pulitzer Prize in Explanatory Reporting in 2011 and won a Gerald Loeb Award in 2010.



Jeff Larson is the Data Editor at ProPublica. He is a winner of the Livingston Award for the 2011 series **Redistricting: How Powerful Interests are Drawing You Out of a Vote**. Jeff's public key can be found [here](#).

Bias in Computer Systems

BATYA FRIEDMAN

Colby College and The Mina Institute
and

HELEN NISSENBAUM

Princeton University

From an analysis of actual cases, three categories of bias in computer systems have been developed: preexisting, technical, and emergent. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. Although others have pointed to bias in particular computer systems and have noted the general problem, we know of no comparable work that examines this phenomenon comprehensively and which offers a framework for understanding and remedying it. We conclude by suggesting that freedom from bias should be counted among the select set of criteria—including reliability, accuracy, and efficiency—according to which the quality of systems in use in society should be judged.

Categories and Subject Descriptors: D.2.0 [**Software**]: Software Engineering; H.1.2 [**Information Systems**]: User/Machine Systems; K.4.0 [**Computers and Society**]: General

General Terms: Design, Human Factors

Additional Key Words and Phrases: Bias, computer ethics, computers and society, design methods, ethics, human values, standards, social computing, social impact, system design, universal design, values

INTRODUCTION

To introduce what bias in computer systems might look like, consider the case of computerized airline reservation systems, which are used widely by travel agents to identify and reserve airline flights for their customers. These reservation systems seem straightforward. When a travel agent types in a customer's travel requirements, the reservation system searches

This research was funded in part by the Clare Boothe Luce Foundation.

Earlier aspects of this work were presented at the 4S/EASST Conference, Goteborg, Sweden, August 1992, and at InterCHI '93, Amsterdam, April 1993. An earlier version of this article appeared as Tech. Rep. CSLI-94-188, CSLI, Stanford University.

Authors' addresses: B. Friedman, Department of Mathematics and Computer Science, Colby College, Waterville, ME 04901; email: b_friedm@colby.edu; H. Nissenbaum, University Center for Human Values, Marx Hall, Princeton University, Princeton, NJ 08544; email: helen@phoenix.princeton.edu. Permission to make digital/hard copy of part or all of this work for personal or classroom use is granted without fee provided that the copies are not made or distributed for profit or commercial advantage, the copyright notice, the title of the publication, and its date appear, and notice is given that copying is by permission of the ACM, Inc. To copy otherwise, to republish, to post on servers, or to redistribute to lists, requires prior specific permission and/or a fee.

© 1996 ACM 1046-8188/96/0700-0330 \$03.50

a database of flights and retrieves all reasonable flight options that meet or come close to the customer's requirements. These options then are ranked according to various criteria, giving priority to nonstop flights, more direct routes, and minimal total travel time. The ranked flight options are displayed for the travel agent. In the 1980s, however, most of the airlines brought before the Antitrust Division of the United States Justice Department allegations of anticompetitive practices by American and United Airlines whose reservation systems—Sabre and Apollo, respectively—dominated the field. It was claimed, among other things, that the two reservations systems are biased [Schrifin 1985].

One source of this alleged bias lies in Sabre's and Apollo's algorithms for controlling search and display functions. In the algorithms, preference is given to "on-line" flights, that is, flights with all segments on a single carrier. Imagine, then, a traveler who originates in Phoenix and flies the first segment of a round-trip overseas journey to London on American Airlines, changing planes in New York. All other things being equal, the British Airlines' flight from New York to London would be ranked lower than the American Airlines' flight from New York to London even though in both cases a traveler is similarly inconvenienced by changing planes and checking through customs. Thus, the computer systems systematically downgrade and, hence, are biased against international carriers who fly few, if any, internal U.S. flights, and against internal carriers who do not fly international flights [Fotos 1988; Ott 1988].

Critics also have been concerned with two other problems. One is that the interface design compounds the bias in the reservation systems. Lists of ranked flight options are displayed screen by screen. Each screen displays only two to five options. The advantage to a carrier of having its flights shown on the first screen is enormous since 90% of the tickets booked by travel agents are booked by the first screen display [Taib 1990]. Even if the biased algorithm and interface give only a small percent advantage overall to one airline, it can make the difference to its competitors between survival and bankruptcy. A second problem arises from the travelers' perspective. When travelers contract with an independent third party—a travel agent—to determine travel plans, travelers have good reason to assume they are being informed accurately of their travel options; in many situations, that does not happen.

As Sabre and Apollo illustrate, biases in computer systems can be difficult to identify let alone remedy because of the way the technology engages and extenuates them. Computer systems, for instance, are comparatively inexpensive to disseminate, and thus, once developed, a biased system has the potential for widespread impact. If the system becomes a standard in the field, the bias becomes pervasive. If the system is complex, and most are, biases can remain hidden in the code, difficult to pinpoint or explicate, and not necessarily disclosed to users or their clients. Furthermore, unlike in our dealings with biased individuals with whom a potential victim can negotiate, biased systems offer no equivalent means for appeal.

Although others have pointed to bias in particular computer systems and have noted the general problem [Johnson and Mulvey 1993; Moor 1985], we know of no comparable work that focuses exclusively on this phenomenon and examines it comprehensively.

In this article, we provide a framework for understanding bias in computer systems. From an analysis of actual computer systems, we have developed three categories: preexisting bias, technical bias, and emergent bias. Preexisting bias has its roots in social institutions, practices, and attitudes. Technical bias arises from technical constraints or considerations. Emergent bias arises in a context of use. We begin by defining bias and explicating each category and then move to case studies. We conclude with remarks about how bias in computer systems can be remedied.

1. WHAT IS A BIASED COMPUTER SYSTEM?

In its most general sense, the term bias means simply “slant.” Given this undifferentiated usage, at times the term is applied with relatively neutral content. A grocery shopper, for example, can be “biased” by not buying damaged fruit. At other times, the term bias is applied with significant moral meaning. An employer, for example, can be “biased” by refusing to hire minorities. In this article we focus on instances of the latter, for if one wants to develop criteria for judging the quality of systems in use—which we do—then criteria must be delineated in ways that speak robustly yet precisely to relevant social matters. Focusing on bias of moral import does just that.

Accordingly, we use the term bias to refer to computer systems that *systematically* and *unfairly discriminate* against certain individuals or groups of individuals in favor of others. A system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate. Consider, for example, an automated credit advisor that assists in the decision of whether or not to extend credit to a particular applicant. If the advisor denies credit to individuals with consistently poor payment records we do not judge the system to be biased because it is reasonable and appropriate for a credit company to want to avoid extending credit privileges to people who consistently do not pay their bills. In contrast, a credit advisor that systematically assigns poor credit ratings to individuals with ethnic surnames discriminates on grounds that are not relevant to credit assessments and, hence, discriminates unfairly.

Two points follow. First, unfair discrimination alone does not give rise to bias unless it occurs systematically. Consider again the automated credit advisor. Imagine a random glitch in the system which changes in an isolated case information in a copy of the credit record for an applicant who happens to have an ethnic surname. The change in information causes a downgrading of this applicant’s rating. While this applicant experiences unfair discrimination resulting from this random glitch, the applicant could have been anybody. In a repeat incident, the same applicant or others with

similar ethnicity would not be in a special position to be singled out. Thus, while the system is prone to random error, it is not biased.

Second, systematic discrimination does not establish bias unless it is joined with an unfair outcome. A case in point is the Persian Gulf War, where United States Patriot missiles were used to detect and intercept Iraqi Scud missiles. At least one software error identified during the war contributed to systematically poor performance by the Patriots [Gao 1992]. Calculations used to predict the location of a Scud depended in complex ways on the Patriots' internal clock. The longer the Patriot's continuous running time, the greater the imprecision in the calculation. The deaths of at least 28 Americans in Dhahran can be traced to this software error, which systematically degraded the accuracy of Patriot missiles. While we are not minimizing the serious consequence of this systematic computer error, it falls outside of our analysis because it does not involve unfairness.

2. FRAMEWORK FOR ANALYZING BIAS IN COMPUTER SYSTEMS

We derived our framework by examining actual computer systems for bias. Instances of bias were identified and characterized according to their source, and then the characterizations were generalized to more abstract categories. These categories were further refined by their application to other instances of bias in the same or additional computer systems. In most cases, our knowledge of particular systems came from the published literature. In total, we examined 17 computer systems from diverse fields including banking, commerce, computer science, education, medicine, and law.

The framework that emerged from this methodology is comprised of three overarching categories—preexisting bias, technical bias, and emergent bias. Table I contains a detailed description of each category. In more general terms, they can be described as follows.

2.1 Preexisting Bias

Preexisting bias has its roots in social institutions, practices, and attitudes. When computer systems embody biases that exist independently, and usually prior to the creation of the system, then we say that the system embodies preexisting bias. Preexisting biases may originate in society at large, in subcultures, and in formal or informal, private or public organizations and institutions. They can also reflect the personal biases of individuals who have significant input into the design of the system, such as the client or system designer. This type of bias can enter a system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions. For example, imagine an expert system that advises on loan applications. In determining an applicant's credit risk, the automated loan advisor negatively weights applicants who live in "undesirable" locations, such as low-income or high-crime neighborhoods, as indicated by their home addresses (a practice referred to as "red-lining"). To the extent the program

Table I. Categories of Bias in Computer System Design

These categories describe ways in which bias can arise in the design of computer systems. The illustrative examples portray plausible cases of bias.

1. Preexisting Bias

Preexisting bias has its roots in social institutions, practices, and attitudes.

When computer systems embody biases that exist independently, and usually prior to the creation of the system, then the system exemplifies preexisting bias. Preexisting bias can enter a system either through the explicit and conscious efforts of individuals or institutions, or implicitly and unconsciously, even in spite of the best of intentions.

1.1. Individual

Bias that originates from individuals who have significant input into the design of the system, such as the client commissioning the design or the system designer (e.g., a client embeds personal racial biases into the specifications for loan approval software).

1.2 Societal

Bias that originates from society at large, such as from organizations (e.g., industry), institutions (e.g., legal systems), or culture at large (e.g., gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls).

2. Technical Bias

Technical bias arises from technical constraints or technical considerations.

2.1 Computer Tools

Bias that originates from a limitation of the computer technology including hardware, software, and peripherals (e.g., in a database for matching organ donors with potential transplant recipients certain individuals retrieved and displayed on initial screens are favored systematically for a match over individuals displayed on later screens).

2.2 Decontextualized Algorithms

Bias that originates from the use of an algorithm that fails to treat all groups fairly under all significant conditions (e.g., a scheduling algorithm that schedules airplanes for take-off relies on the alphabetic listing of the airlines to rank order flights ready within a given period of time).

2.3 Random Number Generation

Bias that originates from imperfections in pseudorandom number generation or in the misuse of pseudorandom numbers (e.g., an imperfection in a random-number generator used to select recipients for a scarce drug leads systematically to favoring individuals toward the end of the database).

2.4 Formalization of Human Constructs

Bias that originates from attempts to make human constructs such as discourse, judgments, or intuitions amenable to computers: when we quantify the qualitative, discretize the continuous, or formalize the nonformal (e.g., a legal expert system advises defendants on whether or not to plea bargain by assuming that law can be spelled out in an unambiguous manner that is not subject to human and humane interpretations in context).

Table I. *Continued*

These categories describe ways in which bias can arise in the design of computer systems. The illustrative examples portray plausible cases of bias.

3. Emergent Bias

Emergent bias arises in a context of use with real users. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.

3.1 New Societal Knowledge

Bias that originates from the emergence of new knowledge in society that cannot be or is not incorporated into the system design (e.g., a medical expert system for AIDS patients has no mechanism for incorporating cutting-edge medical discoveries that affect how individuals with certain symptoms should be treated).

3.2 Mismatch between Users and System Design

Bias that originates when the population using the system differs on some significant dimension from the population assumed as users in the design.

3.2.1 Different Expertise

Bias that originates when the system is used by a population with a different knowledge base from that assumed in the design (e.g., an ATM with an interface that makes extensive use of written instructions—"place the card, magnetic tape side down, in the slot to your left"—is installed in a neighborhood with primarily a nonliterate population).

3.2.2 Different Values

Bias that originates when the system is used by a population with different values than those assumed in the design (e.g., educational software to teach mathematics concepts is embedded in a game situation that rewards individualistic and competitive strategies, but is used by students with a cultural background that largely eschews competition and instead promotes cooperative endeavors).

embeds the biases of clients or designers who seek to avoid certain applicants on the basis of group stereotypes, the automated loan advisor's bias is preexisting.

2.2 Technical Bias

In contrast to preexisting bias, technical bias arises from the resolution of issues in the technical design. Sources of technical bias can be found in several aspects of the design process, including limitations of computer tools such as hardware, software, and peripherals; the process of ascribing social meaning to algorithms developed out of context; imperfections in pseudorandom number generation; and the attempt to make human constructs amenable to computers, when we quantify the qualitative, discretize the continuous, or formalize the nonformal. As an illustration, consider again the case of Sabre and Apollo described above. A technical constraint imposed by the size of the monitor screen forces a piecemeal presentation of flight options and, thus, makes the algorithm chosen to

rank flight options critically important. Whatever ranking algorithm is used, if it systematically places certain airlines' flights on initial screens and other airlines' flights on later screens, the system will exhibit technical bias.

2.3 Emergent Bias

While it is almost always possible to identify preexisting bias and technical bias in a system design at the time of creation or implementation, emergent bias arises only in a context of use. This bias typically emerges some time after a design is completed, as a result of changing societal knowledge, population, or cultural values. Using the example of an automated airline reservation system, envision a hypothetical system designed for a group of airlines all of whom serve national routes. Consider what might occur if that system was extended to include international airlines. A flight-ranking algorithm that favors on-line flights when applied in the original context with national airlines leads to no systematic unfairness. However, in the new context with international airlines, the automated system would place these airlines at a disadvantage and, thus, comprise a case of emergent bias. User interfaces are likely to be particularly prone to emergent bias because interfaces by design seek to reflect the capacities, character, and habits of prospective users. Thus, a shift in context of use may well create difficulties for a new set of users.

3. APPLICATIONS OF THE FRAMEWORK

We now analyze actual computer systems in terms of the framework introduced above. It should be understood that the systems we analyze are by and large good ones, and our intention is not to undermine their integrity. Rather, our intention is to develop the framework, show how it can identify and clarify our understanding of bias in computer systems, and establish its robustness through real-world cases.

3.1 The National Resident Match Program (NRMP)

The NRMP implements a centralized method for assigning medical school graduates their first employment following graduation. The centralized method of assigning medical students to hospital programs arose in the 1950s in response to the chaotic job placement process and on-going failure of hospitals and students to arrive at optimal placements. During this early period the matching was carried out by a mechanical card-sorting process, but in 1974 electronic data processing was introduced to handle the entire matching process. (For a history of the NRMP, see Graettinger and Peranson [1981a].) After reviewing applications and interviewing students, hospital programs submit to the centralized program their ranked list of students. Students do the same for hospital programs. Hospitals and students are not permitted to make other arrangements with one another or to attempt to directly influence each others' rankings prior to the match.

Algorithmic Fairness

review articles

DOI:10.1145/3376898

A group of industry, academic, and government experts convene in Philadelphia to explore the roots of algorithmic bias.

BY ALEXANDRA CHOULDECHOVA AND AARON ROTH

A Snapshot of the Frontiers of Fairness in Machine Learning

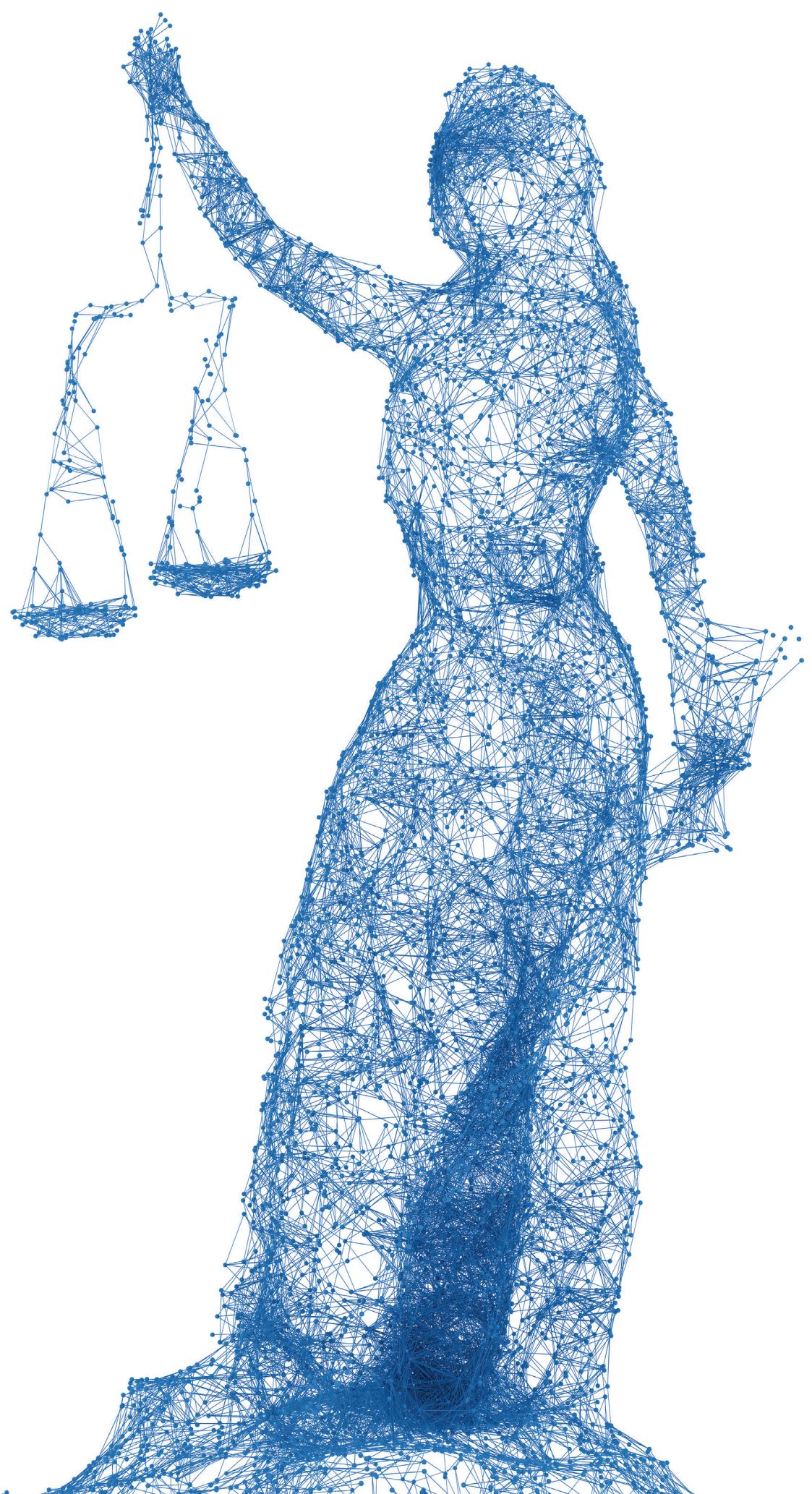
THE LAST DECADE has seen a vast increase both in the diversity of applications to which machine learning is applied, and to the import of those applications. Machine learning is no longer just the engine behind ad placements and spam filters; it is now used to filter loan applicants, deploy police officers, and inform bail and parole decisions, among other things. The result has been a major concern for the potential for data-driven methods to introduce and perpetuate discriminatory practices, and to otherwise be unfair. And this concern has not been without reason: a steady stream of empirical findings has shown that data-driven methods can unintentionally both encode existing human biases and introduce new ones.^{7,9,11,60}

At the same time, the last two years have seen an unprecedented explosion in interest from the academic community in studying fairness and machine learning. “Fairness and transparency” transformed from a niche topic with a trickle of papers produced every year (at least since the work of Pedresh⁵⁶) to a major subfield of machine learning, complete with a dedicated archival conference—ACM FAT*. But despite the volume and velocity of published work, our understanding of the fundamental questions related to fairness and machine learning remain in its infancy. What should fairness mean? What are the causes that introduce unfairness in machine learning? How best should we modify our algorithms to avoid unfairness? And what are the corresponding trade offs with which we must grapple?

In March 2018, we convened a group of about 50 experts in Philadelphia, drawn from academia, industry, and government, to assess the state of our understanding of the fundamentals of the nascent science of fairness in machine learning, and to identify the unanswered questions that seem the most pressing. By necessity, the aim of the workshop was not to comprehensively cover the vast growing field, much of which is empirical. Instead, the focus was on theoretical work aimed at providing a scientific foundation for understanding algo-

» key insights

- The algorithmic fairness literature is enormous and growing quickly, but our understanding of basic questions remains nascent.
- Researchers have yet to find entirely compelling definitions, and current work focuses mostly on supervised learning in static settings.
- There are many compelling open questions related to robustly accounting for the effects of interventions in dynamic settings, learning in the presence of data contaminated with human bias, and finding definitions of fairness that guarantee individual-level semantics while remaining actionable.



rithmic bias. This document captures several of the key ideas and directions discussed. It is not an exhaustive account of work in the area.

What We Know

Even before we precisely specify what we mean by “fairness,” we can identify common distortions that can lead off-the-shelf machine learning techniques to produce behavior that is intuitively unfair. These include:

1. *Bias encoded in data.* Often, the training data we have on hand already includes human biases. For example, in the problem of recidivism prediction used to inform bail and parole decisions, the goal is to predict whether an inmate, if released, will go on to commit another crime within a fixed period of time. But we do not have data on who commits crimes—we have data on who is arrested. There is reason to believe that arrest data—especially for drug crimes—is skewed toward minority populations that are policed at a higher rate.⁵⁹ Of course, machine learning techniques are designed to fit the data, and so will naturally replicate any bias already present in the data. There is no reason to expect them to remove existing bias.

2. *Minimizing average error fits majority populations.* Different populations of people have different distributions over features, and those features have different relationships to the label that we are trying to predict. As an example, consider the task of predicting college performance based on high school data. Suppose there is a majority population and a minority population. The majority population employs SAT tutors and takes the exam multiple times, reporting only the highest score. The minority population does not. We should naturally expect both that SAT scores are higher among the majority population, and that their relationship to college performance is differently calibrated compared to the minority population. But if we train a group-blind classifier to minimize overall error, if it cannot simultaneously fit both populations optimally, it will fit the majority population. This is because—simply by virtue of their numbers—the fit to the majority population is more important to overall error than the fit to

Given the limitations of extant notions of fairness, is there a way to get some of the “best of both worlds?”

the minority population. This leads to a different (and higher) distribution of errors in the minority population. This effect can be quantified and can be partially alleviated via concerted data gathering effort.¹⁴

3. *The need to explore.* In many important problems, including recidivism prediction and drug trials, the data fed into the prediction algorithm depends on the actions that algorithm has taken in the past. We only observe whether an inmate will recidivate if we release him. We only observe the efficacy of a drug on patients to whom it is assigned. Learning theory tells us that in order to effectively learn in such scenarios, we need to explore—that is, sometimes take actions we believe to be sub-optimal in order to gather more data. This leads to at least two distinct ethical questions. First, when are the individual costs of exploration borne disproportionately by a certain sub-population? Second, if in certain (for example, medical) scenarios, we view it as immoral to take actions we believe to be sub-optimal for any particular patient, how much does this slow learning, and does this lead to other sorts of unfairness?

Definitions of fairness. With a few exceptions, the vast majority of work to date on fairness in machine learning has focused on the task of batch classification. At a high level, this literature has focused on two main families of definitions:^a statistical notions of fairness and individual notions of fairness. We briefly review what is known about these approaches to fairness, their advantages, and their shortcomings.

Statistical definitions of fairness. Most of the literature on fair classification focuses on statistical definitions of fairness. This family of definitions fixes a small number of protected demographic groups G (such as racial groups), and then ask for (approximate) parity of some statistical measure across all of these groups. Popular measures include raw positive classification rate, considered in

^a There is also an emerging line of work that considers causal notions of fairness (for example, see Kilbertus,⁴³ Kusner,⁴⁸ Nabi⁵⁵). We intentionally avoided discussions of this potentially important direction because it will be the subject of its own CCC visioning workshop.

work such as Calders,¹⁰ Dwork,¹⁹ Feldman,²⁵ Kamishima,³⁶ (also sometimes known as statistical parity,¹⁹ false positive and false negative rates^{15,29,46,63} (also sometimes known as equalized odds²⁹), and positive predictive value^{15,46} (closely related to equalized calibration when working with real valued risk scores). There are others—see, for example, Berk⁴ for a more exhaustive enumeration.

This family of fairness definitions is attractive because it is simple, and definitions from this family can be achieved without making any assumptions on the data and can be easily verified. However, statistical definitions of fairness do not on their own give meaningful guarantees to individuals or structured subgroups of the protected demographic groups. Instead they give guarantees to “average” members of the protected groups. (See Dwork¹⁹ for a litany of ways in which statistical parity and similar notions can fail to provide meaningful guarantees, and Kearns⁴⁰ for examples of how some of these weaknesses carry over to definitions that equalize false positive and negative rates.) Different statistical measures of fairness can be at odds with one another. For example, Chouldechova¹⁵ and Kleinberg⁴⁶ prove a fundamental impossibility result: except in trivial settings, it is impossible to simultaneously equalize false positive rates, false negative rates, and positive predictive value across protected groups. Learning subject to statistical fairness constraints can also be computationally hard,⁶¹ although practical algorithms of various sorts are known.^{1,29,63}

Individual definitions of fairness. Individual notions of fairness, on the other hand, ask for constraints that bind on specific pairs of individuals, rather than on a quantity that is averaged over groups. For example, Dwork¹⁹ gives a definition which roughly corresponds to the constraint that “similar individuals should be treated similarly,” where similarity is defined with respect to a task-specific metric that must be determined on a case by case basis. Joseph³⁵ suggests a definition that corresponds approximately to “less qualified individuals should not be favored over more qualified individuals,” where quality is de-

fined with respect to the true underlying label (unknown to the algorithm). However, although the semantics of these kinds of definitions can be more meaningful than statistical approaches to fairness, the major stumbling block is that they seem to require making significant assumptions. For example, the approach of Dwork¹⁹ presupposes the existence of an agreed upon similarity metric, whose definition would itself seemingly require solving a non-trivial problem in fairness, and the approach of Joseph³⁵ seems to require strong assumptions on the functional form of the relationship between features and labels in order to be usefully put into practice. These obstacles are serious enough that it remains unclear whether individual notions of fairness can be made practical—although attempting to bridge this gap is an important and ongoing research agenda.

Questions at the Research Frontier

Given the limitations of extant notions of fairness, is there a way to get some of the “best of both worlds?” In other words, constraints that are practically implementable without the need for making strong assumptions on the data or the knowledge of the algorithm designer, but which nevertheless provide more meaningful guarantees to individuals? Two recent papers, Kearns⁴⁰ and Hèbert-Johnson³⁰ (see also Kearns⁴² and Kim⁴⁴ for empirical evaluations of the algorithms proposed in these papers), attempt to do this by asking for statistical fairness definitions to hold not just on a small number of protected groups, but on an exponential or infinite class of groups defined by some class of functions of bounded complexity. This approach seems promising—because, ultimately, they are asking for statistical notions of fairness—the approaches proposed by these papers enjoy the benefits of statistical fairness: that no assumptions need be made about the data, nor is any external knowledge (like a fairness metric) needed. It also better addresses concerns about “intersectionality,” a term used to describe how different kinds of discrimination can compound and interact for individuals who fall at the intersection of

several protected classes.

At the same time, the approach raises a number of additional questions: What function classes are reasonable, and once one is decided upon (for example, conjunctions of protected attributes), what features should be “protected?” Should these only be attributes that are sensitive on their own, like race and gender, or might attributes that are innocuous on their own correspond to groups we wish to protect once we consider their intersection with protected attributes (for example clothing styles intersected with race or gender)? Finally, this family of approaches significantly mitigates some of the weaknesses of statistical notions of fairness by asking for the constraints to hold on average not just over a small number of coarsely defined groups, but over very finely defined groups as well. Ultimately, however, it inherits the weaknesses of statistical fairness as well, just on a more limited scale.

Another recent line of work aims to weaken the strongest assumption needed for the notion of individual fairness from Dwork:¹⁹ namely the algorithm designer has perfect knowledge of a “fairness metric.” Kim⁴⁵ assumes the algorithm has access to an oracle which can return an unbiased estimator for the distance between two randomly drawn individuals according to an unknown fairness metric, and show how to use this to ensure a statistical notion of fairness related to Hèbert-Johnson³⁰ and Kearns,⁴⁰ which informally state that “on average, individuals in two groups should be treated similarly if on average the individuals in the two groups are similar” and this can be achieved with respect to an exponentially or infinitely large set of groups. Similarly, Gillen²⁸ assumes the existence of an oracle, which can identify fairness violations when they are made in an online setting but cannot quantify the extent of the violation (with respect to the unknown metric). It is shown that when the metric is from a specific learnable family, this kind of feedback is sufficient to obtain an optimal regret bound to the best fair classifier while having only a bounded number of violations of the fairness metric. Rothblum⁵⁸ considers the case in which

the metric is known and show that a PAC-inspired approximate variant of metric fairness generalizes to new data drawn from the same underlying distribution. Ultimately, however, these approaches all assume fairness is perfectly defined with respect to some metric, and that there is some sort of direct access to it. Can these approaches be generalized to a more “agnostic” setting, in which fairness feedback is given by human beings who may not be responding in a way that is consistent with any metric?

Data evolution and dynamics of fairness. The vast majority of work in computer science on algorithmic fairness has focused on one-shot classification tasks. But real algorithmic systems consist of many different components combined together, and operate in complex environments that are dynamically changing, sometimes because of the actions of the learning algorithm itself. For the field to progress, we need to understand the dynamics of fairness in more complex systems.

Perhaps the simplest aspect of dynamics that remains poorly understood is how and when components that may individually satisfy notions of fairness compose into larger constructs that still satisfy fairness guarantees. For example, if the bidders in an advertising auction individually are fair with respect to their bidding decisions, when will the allocation of advertisements be fair, and when will it not? Bower⁸ and Dwork²⁰ have made a preliminary foray in this direction. These papers embark on a systematic study of fairness under composition and find that often the composition of multiple fair components will not satisfy any fairness constraint at all. Similarly, the individual components of a fair system may appear to be unfair in isolation. There are certain special settings, for example, the “filtering pipeline” scenario of Bower⁸—modeling a scenario in which a job applicant is selected only if she is selected at every stage of the pipeline—in which (multiplicative approximations of) statistical fairness notions compose in a well behaved way. But the high-level message from these works is that our current notions of fairness compose poorly. Experience

from differential privacy^{21,22} suggests that graceful degradation under composition is key to designing complicated algorithms satisfying desirable statistical properties, because it allows algorithm design and analysis to be modular. Thus, it seems important to find satisfying fairness definitions and richer frameworks that behave well under composition.

In dealing with socio-technical systems, it is also important to understand how algorithms dynamically effect their environment, and the incentives of human actors. For example, if the bar (for example, college admission) is lowered for a group of individuals, this might increase the average qualifications for this group over time because of at least two effects: a larger proportion of children in the next generation grow up in households with college educated parents (and the opportunities this provides), and the fact that a college education is achievable can incentivize effort to prepare academically. These kinds of effects are not considered when considering either statistical or individual notions of fairness in one-shot learning settings.

The economics literature on affirmative action has long considered such effects—although not with the specifics of machine learning in mind: see, for example, Becker,³ Coat,¹⁶ Foster.²⁶ More recently, there have been some preliminary attempts to model these kinds of effects in machine learning settings—for example, by modeling the environment as a Markov decision process,³² considering the equilibrium effects of imposing statistical definitions of fairness in a model of a labor market,³¹ specifying the functional relationship between classification outcomes and quality,⁴⁹ or by considering the effect of a classifier on a downstream Bayesian decision maker.³⁹ However, the specific predictions of most of the models of this sort are brittle to the specific modeling assumptions made—they point to the need to consider long term dynamics, but do not provide robust guidance for how to navigate them. More work is needed here.

Finally, decision making is often distributed between a large number of actors who share different goals

and do not necessarily coordinate. In settings like this, in which we do not have direct control over the decision-making process, it is important to think about how to incentivize rational agents to behave in a way that we view as fair. Kannan³⁷ takes a preliminary stab at this task, showing how to incentivize a particular notion of individual fairness in a simple, stylized setting, using small monetary payments. But how should this work for other notions of fairness, and in more complex settings? Can this be done by controlling the flow of information, rather than by making monetary payments (monetary payments might be distasteful in various fairness-relevant settings)? More work is needed here as well. Finally, Corbett-Davies¹⁷ take a welfare maximization view of fairness in classification and characterize the cost of imposing additional statistical fairness constraints as well. But this is done in a static environment. How would the conclusions change under a dynamic model?

Modeling and correcting bias in the data. Fairness concerns typically surface precisely in settings where the available training data is already contaminated by bias. The data itself is often a product of social and historical process that operated to the disadvantage of certain groups. When trained in such data, off-the-shelf machine learning techniques may reproduce, reinforce, and potentially exacerbate existing biases. Understanding how bias arises in the data, and how to correct for it, are fundamental challenges in the study of fairness in machine learning.

Bolukbasi⁷ demonstrate how machine learning can reproduce biases in their analysis of the popular word2vec embedding trained on a corpus of Google News texts (parallel effects were independently discovered by Caliskan¹¹). The authors show that the trained embedding exhibit female/male gender stereotypes, learning that “doctor” is more similar to man than to woman, along with analogies such as “man is to computer programmer as woman is to homemaker.” Even if such learned associations accurately reflect patterns in the source text corpus, their use in automated systems may exacerbate existing bi-

ases. For instance, it might result in male applicants being ranked more highly than equally qualified female applicants in queries related to jobs that the embedding identifies as male-associated.

Similar risks arise whenever there is potential for feedback loops. These are situations where the trained machine learning model informs decisions that then affect the data collected for future iterations of the training process. Lum⁵¹ demonstrate how feedback loops might arise in predictive policing if arrest data were used to train the model.^b In a nutshell, since police are likely to make more arrests in more heavily policed areas, using arrest data to predict crime hotspots will disproportionately concentrate policing efforts on already over-policed communities. Expanding on this analysis, Ensign²⁴ finds that incorporating community-driven data, such as crime reporting, helps to attenuate the biasing feedback effects. The authors also propose a strategy for accounting for feedback by adjusting arrest counts for policing intensity. The success of the mitigation strategy, of course, depends on how well the simple theoretical model reflects the true relationships between crime intensity, policing, and arrests. Problematically, such relationships are often unknown, and are very difficult to infer from data. This situation is by no means specific to predictive policing.

Correcting for data bias generally seems to require knowledge of how the measurement process is biased, or judgments about properties the data would satisfy in an “unbiased” world. Friedler²⁷ formalize this as a disconnect between the *observed space*—features that are observed in the data, such as SAT scores—and the unobservable *construct space*—features that form the desired basis for decision making, such as intelligence. Within this framework, data correction efforts attempt to undo the effects of biasing mechanisms that drive discrepancies between these spaces. To the extent that the biasing

Fairness concerns typically surface precisely in settings where the available training data is already contaminated by bias.

mechanism cannot be inferred empirically, any correction effort must make explicit its underlying assumptions about this mechanism. What precisely is being assumed about the construct space? When can the mapping between the construct space and the observed space be learned and inverted? What form of fairness does the correction promote, and at what cost? The costs are often immediately realized, whereas the benefits are less tangible. We will directly observe reductions in prediction accuracy, but any gains hinge on a belief that the observed world is not one we should seek to replicate accurately in the first place. This is an area where tools from causality may offer a principled approach for drawing valid inference with respect to unobserved counterfactually ‘fair’ worlds.

Fair representations. Fair representation learning is a data debiasing process that produces transformations (intermediate representations) of the original data that retain as much of the task-relevant information as possible while removing information about sensitive or protected attributes. This is one approach to transforming biased observational data in which group membership may be inferred from other features, to a construct space where protected attributes are statistically independent of other features.

First introduced in the work of Zemel⁶⁴ fair representation learning produces a debiased data set that may in principle be used by other parties without any risk of disparate outcomes. Feldman²⁵ and McNamara⁵⁴ formalize this idea by showing how the disparate impact of a decision rule is bounded in terms of its balanced error rate as a predictor of the sensitive attribute.

Several recent papers have introduced new approaches for constructing fair representations. Feldman²⁵ propose rank-preserving procedures for repairing features to reduce or remove pairwise dependence with the protected attribute. Johndrow³³ build upon this work, introducing a likelihood-based approach that can additionally handle continuous protected attributes, discrete features, and which promotes joint independence

^b Predictive policing models are generally proprietary, and so it is not clear whether arrest data is used to train the model in any deployed system.

between the transformed features and the protected attributes. There is also a growing literature on using adversarial learning to achieve group fairness in the form of statistical parity or false positive/false negative rate balance.^{5,23,52,65}

Existing theory shows the fairness-promoting benefits of fair-representation learning rely critically on the extent to which existing associations between the transformed features and the protected characteristics are removed. Adversarial downstream users may be able to recover protected attribute information if their models are more powerful than those used initially to obfuscate the data. This presents a challenge both to the generators of fair representations as well as to auditors and regulators tasked with certifying that the resulting data is fair for use. More work is needed to understand the implications of fair representation learning for promoting fairness in the real world.

Beyond classification. Although the majority of the work on fairness in machine learning focuses on batch classification, it is but one aspect of how machine learning is used. Much of machine learning—for example, online learning, bandit learning, and reinforcement learning—focuses on dynamic settings in which the actions of the algorithm feed back into the data it observes. These dynamic settings capture many problems for which fairness is a concern. For example, lending, criminal recidivism prediction, and sequential drug trials are so-called bandit learning problems, in which the algorithm cannot observe data corresponding to counterfactuals. We cannot see whether someone not granted a loan would have paid it back. We cannot see whether an inmate not released on parole would have gone on to commit another crime. We cannot see how a patient would have responded to a different drug.

The theory of learning in bandit settings is well understood, and it is characterized by a need to trade-off exploration with exploitation. Rather than always making a myopically optimal decision, when counterfactuals cannot be observed, it is necessary for algorithms to sometimes take ac-

Much of machine learning focuses on dynamic settings in which the actions of the algorithm feed back into the data it observes. These dynamic settings capture many problems for which fairness is a concern.

tions that appear to be sub-optimal so as to gather more data. But in settings in which decisions correspond to individuals, this means sacrificing the well-being of a particular person for the potential benefit of future individuals. This can sometimes be unethical, and a source of unfairness.⁶ Several recent papers explore this issue. For example, Bastani² and Kannan³⁸ give conditions under which linear learners need not explore at all in bandit settings, thereby allowing for best-effort service to each arriving individual, obviating the tension between ethical treatment of individuals and learning. Raghavan⁵⁷ show the costs associated with exploration can be unfairly born by a structured sub-population, and that counter-intuitively, those costs can actually increase when they are included with a majority population, even though more data increases the rate of learning overall. However, these results are all preliminary: they are restricted to settings in which the learner is learning a linear policy, and the data really is governed by a linear model. While illustrative, more work is needed to understand real-world learning in online settings, and the ethics of exploration.

There is also some work on fairness in machine learning in other settings—for example, ranking,¹² selection,^{42,47} personalization,¹³ bandit learning,^{34,50} human-classifier hybrid decision systems,⁵³ and reinforcement learning.^{18,32} But outside of classification, the literature is relatively sparse. This should be rectified, because there are interesting and important fairness issues that arise in other settings—especially when there are combinatorial constraints on the set of individuals that can be selected for a task, or when there is a temporal aspect to learning.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 1136993. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

We are indebted to all of the participants of the CCC visioning work-

shop; discussions from that meeting shaped every aspect of this document. Also, our thanks to Helen Wright, Ann Drobnić, Cynthia Dwork, Sampath Kannan, Michael Kearns, Toni Pitassi, and Suresh Venkatasubramanian. □

References

- Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J. and Wallach, H. A reductions approach to fair classification. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, JMLR Workshop and Conference Proceedings, 2018, 2569–2577.
- Bastani, H., Bayati, M. and Khosravi, K. Exploiting the natural exploration in contextual bandits. arXiv preprint, 2017, arXiv:1704.09011.
- Becker, G.S. *The Economics of Discrimination*. University of Chicago Press, 2010.
- Berk, R., Heidari, H., Jabbari, S., Kearns, M. and Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0(0):004912411872533.
- Beutel, A., Chen, J., Zhao, Z. and Chi, E.H. Data decisions and theoretical implications when adversarially learning fair representations. arXiv preprint, 2017, arXiv:1707.00075.
- Bird, S., Barocas, S., Crawford, K., Diaz, F. and Wallach, H. Exploring or exploiting? Social and ethical implications of autonomous experimentation in AI. In *Proceedings of Workshop on Fairness, Accountability, and Transparency in Machine Learning*. ACM, 2016.
- Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V. and Kalai, A.T. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, 2016, 4349–4357.
- Bower, A. et al. Fair pipelines. arXiv preprint, 2017, arXiv:1707.00391.
- Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the Conference on Fairness, Accountability and Transparency*. ACM, 2018, 77–91.
- Calders, T. and Verwer, S. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (2010), 277–292.
- Caliskan, A., Bryson, J.J. and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
- Celis, L.E., Straszak, D. and Vishnoi, N.K. Ranking with fairness constraints. In *Proceedings of the 45th Intern. Colloquium on Automata, Languages, and Programming*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.
- Celis, L.E. and Vishnoi, N.K. Fair personalization. arXiv preprint, 2017, arXiv:1707.02260.
- Chen, I., Johansson, F.D. and Sontag, D. Why is my classifier discriminatory? *Advances in Neural Information Processing Systems*, 2018, 3539–3550.
- Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
- Coat, S. and Loury, G.C. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, 1993, 1220–1240.
- Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. and Huq, A. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM, 2017, 797–806.
- Doroudi, S., Thomas, P.S. and Brunskill, E. Importance sampling for fair policy selection. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2017.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O. and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conf.* ACM, 2012, 214–226.
- Dwork, C. and Ilvento, C. Fairness under composition. Manuscript, 2018.
- Dwork, C., McSherry, F., Nissim, K. and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Proceedings of Theory of Cryptography Conference*. Springer, 2006, 265–284.
- Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3–4 (2014), 211–407.
- Edwards, H. and Storkey, A. Censoring representations with an adversary. arXiv preprint, 2015, arXiv:1511.05897.
- Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C. and Venkatasubramanian, S. Runaway feedback loops in predictive policing. In *Proceedings of 1st Conf. Fairness, Accountability and Transparency in Computer Science*. ACM, 2018.
- Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. Certifying and removing disparate impact. *Proceedings of KDD*, 2015.
- Foster, D.P. and Vohra, R.A. An economic argument for affirmative action. *Rationality and Society* 4, 2 (1992), 176–188.
- Friedler, S.A., Scheidegger, C. and Venkatasubramanian, S. On the (im) possibility of fairness. arXiv preprint, 2016, arXiv:1609.07236.
- Gillen, S., Jung, C., Kearns, M. and Roth, A. Online learning with an unknown fairness metric. *Advances in Neural Information Processing Systems*, 2018.
- Hardt, M., Price, E. and Srebro, N. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 2016, 3315–3323.
- Hébert-Johnson, U., Kim, M.P., Reingold, O. and Rothblum, G.N. Calibration for the (computationally identifiable) masses. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, JMLR Workshop and Conference Proceedings, 2018, 2569–2577.
- Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. P.A. Champin, F.L. Gandon, M. Lalmas, and P.G. Ipeirotis, eds. ACM, 2018, 1389–1398.
- Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A. Fairness in reinforcement learning. In *Proceedings of the Intern. Conf. Machine Learning*, 2017, 1617–1626.
- Johnsrow, J.E., Lum, K. et al. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics* 13, 1 (2019), 189–220.
- Joseph, M., Kearns, M., Morgenstern, J.H., Neel, S. and Roth, A. Fair algorithms for infinite and contextual bandits. In *Proceedings of AAAI/ACM Conference on AI, Ethics, and Society*, 2018.
- Joseph, M., Kearns, M., Morgenstern, J.H. and Roth, A. Fairness in learning: Classic and contextual bandits. *Advances in Neural Information Processing Systems*, 2016, 325–333.
- Kamishima, T., Akaho, S. and Sakuma, J. Fairness-aware learning through regularization approach. In *Proceedings of the IEEE 11th Intern. Conf. Data Mining Workshops*. IEEE, 2011, 643–650.
- Kannan, S. et al. Fairness incentives for myopic agents. In *Proceedings of the 2017 ACM Conference on Economics and Computation*. ACM, 2017, 369–386.
- Kannan, S., Morgenstern, J., Roth, A., Waggoner, B. and Wu, Z.S. A smoothed analysis of the greedy algorithm for the linear contextual bandit problem. *Advances in Neural Information Processing Systems*, 2018.
- Kannan, S., Roth, A. and Ziani, J. Downstream effects of affirmative action. In *Proceedings of the Conf. Fairness, Accountability, and Transparency*. ACM, 2019, 240–248.
- Kearns, M.J., Neel, S., Roth, A. and Wu, Z.S. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *Proceedings of the 35th International Conference on Machine Learning*. J.G. Dy and A. Krause, eds. JMLR Workshop and Conference Proceedings, ICML, 2018, 2569–2577.
- Kearns, M., Neel, S., Roth, A. and Wu, Z.S. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conf. Fairness, Accountability, and Transparency*. ACM, 2019, 100–109.
- Kearns, M., Roth, A. and Wu, Z.S. Meritocratic fairness for cross-population selection. In *Proceedings of International Conference on Machine Learning*, 2017, 1828–1836.
- Kilbertus, N. et al. Avoiding discrimination through causal reasoning. *Advances in Neural Information Processing Systems*, 2017, 656–666.
- Kim, M.P., Ghorbani, A. and Zou, J. Multiaccuracy: Blackbox postprocessing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, 247–254.
- Kim, M.P., Reingold, O. and Rothblum, G.N. Fairness through computationally bounded awareness. *Advances in Neural Information Processing Systems*, 2018.
- Kleinberg, J.M., Mullainathan, S. and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. In *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017.
- Kleinberg, J. and Raghavan, M. Selection problems in the presence of implicit bias. In *Proceedings of the 9th Innovations in Theoretical Computer Science Conference* 94, 2018, 33. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.
- Kusner, M.J., Loftus, J., Russell, C. and Silva, R. Counterfactual fairness. *Advances in Neural Information Processing Systems*, 2017, 4069–4079.
- Liu, L.T., Dean, S., Rolf, E., Simchowitz, M. and Hardt, M. Delayed impact of fair machine learning. In *Proceedings of the 35th Intern. Conf. Machine Learning*. ICML, 2018.
- Liu, Y., Radanovic, G., Dimitrakakis, C., Mandal, D. and Parkes, D.C. Calibrated fairness in bandits. arXiv preprint, 2017, arXiv:1707.01875.
- Lum, K. and Isaac, W. To predict and serve? *Significance* 13, 5 (2016), 14–19.
- Madras, D., Creager, E., Pitassi, T. and Zemel, R. Learning adversarially fair and transferable representations. In *Proceedings of Intern. Conf. Machine Learning*, 2018, 3381–3390.
- Madras, D., Pitassi, T. and Zemel, R.S. Predict responsibly: Increasing fairness by learning to defer. CoRR, 2017, abs/1711.06664.
- McNamara, D., Ong, C.S. and Williamson, R.C. Provably fair representations. arXiv preprint, 2017, arXiv:1710.04394.
- Nabi, R. and Shpitser, I. Fair inference on outcomes. In *Proceedings of the AAAI Conference on Artificial Intelligence* 2018 (2018), 1931. NIH Public Access.
- Pedreshi, D., Ruggieri, S. and Turini, F. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD Intern. Conf. Knowledge Discovery and Data Mining*. ACM, 2008, 560–568.
- Raghavan, M., Slivkins, A., Wortman Vaughan, J. and Wu, Z.S. The unfair externalities of exploration. *Conference on Learning Theory*, 2018.
- Rothblum, G.N. and Yona, G. Probably approximately metric-fair learning. In *Proceedings of the 35th Intern. Conf. Machine Learning*. JMLR Workshop and Conference Proceedings, ICML 80 (2018), 2569–2577.
- Rothwell, J. How the war on drugs damages black social mobility. The Brookings Institution, Sept. 30, 2014.
- Sweeney, L. Discrimination in online ad delivery. *Queue* 11, 3 (2013), 10.
- Woodworth, B., Gunasekar, S., Ohannessian, M.I. and Srebro, N. Learning non-discriminatory predictors. In *Proceedings of Conf. Learning Theory*, 2017, 1920–1953.
- Yang, K. and Stoyanovich, J. Measuring fairness in ranked outputs. In *Proceedings of the 29th Intern. Conf. Scientific and Statistical Database Management*. ACM, 2017, 22.
- Zafar, M.B., Valera, I., Gomez-Rodriguez, M. and Gummadi, K.P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th Intern. Conf. World Wide Web*. ACM, 2017, 1171–1180.
- Zemel, R., Wu, Y., Swersky, K., Pitassi, T. and Dwork, C. Learning fair representations. In *Proceedings of ICML*, 2013.
- Zhang, B.H., Lemoine, B. and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conf. AI, Ethics, and Society*. ACM, 2018, 335–340.

Alexandra Chouldechova (achould@cmu.edu) is Estella Loomis Assistant Professor of Statistics and Public Policy in the Heinz College at Carnegie Mellon University, Pittsburgh, PA, USA.

Aaron Roth (aaroth@cis.upenn.edu) is Class of 1940 Associate Professor in the Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA. Together with Michael Kearns, he is the author of *The Ethical Algorithm*.

Copyright held by authors/owners.
Publication rights licensed to ACM.



Watch the authors discuss this work in the exclusive *Communications* video.
<https://cacm.acm.org/videos/frontiers-of-fairness>

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg *

Sendhil Mullainathan †

Manish Raghavan ‡

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant’s probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal justice system; in one widely-publicized report, Angwin et al. analyzed a commonly used statistical method for assigning risk scores in the criminal justice system — the COMPAS risk tool — and argued that it was biased against African-American defendants [2, 23]. One of their main contentions was that the tool’s errors were asymmetric: African-American defendants were more likely to be incorrectly labeled as higher-risk than they actually were, while white defendants were more likely to be incorrectly labeled as lower-risk than they actually were. Subsequent analyses raised methodological objections to this report, and also observed that despite the COMPAS risk tool’s errors, its estimates of the probability of recidivism are equally well calibrated to the true outcomes for both African-American and white defendants [1, 10, 13, 17].

*Cornell University

†Harvard University

‡Cornell University

Second, in a very different domain, researchers have begun to analyze the ways in which different genders and racial groups experience advertising and commercial content on the Internet differently [9, 26]. We could ask, for example: if a male user and female user are equally interested in a particular product, does it follow that they’re equally likely to be shown an ad for it? Sometimes this concern may have broader implications, for example if women in aggregate are shown ads for lower-paying jobs. Other times, it may represent a clash with a user’s leisure interests: if a female user interacting with an advertising platform is interested in an activity that tends to have a male-dominated viewership, like professional football, is the platform as likely to show her an ad for football as it is to show such an ad to an interested male user?

A third domain, again quite different from the previous two, is medical testing and diagnosis. Doctors making decisions about a patient’s treatment may rely on tests providing probability estimates for different diseases and conditions. Here too we can ask whether such decision-making is being applied uniformly across different groups of patients [16, 27], and in particular how medical tests may play a differential role for conditions that vary widely in frequency between these groups.

Providing guarantees for decision procedures. One can raise analogous questions in many other domains of fundamental importance, including decisions about hiring, lending, or school admissions [24], but we will focus on the three examples above for the purposes of this discussion. In these three example domains, a few structural commonalities stand out. First, the algorithmic estimates are often being used as “input” to a larger framework that makes the overall decision — a risk score provided to a human expert in the legal and medical instances, and the output of a machine-learning algorithm provided to a larger advertising platform in the case of Internet ads. Second, the underlying task is generally about classifying whether people possess some relevant property: recidivism, a medical condition, or interest in a product. We will refer to people as being *positive instances* if they truly possess the property, and *negative instances* if they do not. Finally, the algorithmic estimates being provided for these questions are generally not pure yes-no decisions, but instead probability estimates about whether people constitute positive or negative instances.

Let us suppose that we are concerned about how our decision procedure might operate differentially between two groups of interest (such as African-American and white defendants, or male and female users of an advertising system). What sorts of guarantees should we ask for as protection against potential bias?

A first basic goal in this literature is that the probability estimates provided by the algorithm should be *well-calibrated*: if the algorithm identifies a set of people as having a probability z of constituting positive instances, then approximately a z fraction of this set should indeed be positive instances [8, 14]. Moreover, this condition should hold when applied separately in each group as well [13]. For example, if we are thinking in terms of potential differences between outcomes for men and women, this means requiring that a z fraction of men and a z fraction of women assigned a probability z should possess the property in question.

A second goal focuses on the people who constitute positive instances (even if the algorithm can only imperfectly recognize them): the average score received by people constituting positive instances should be the same in each group. We could think of this as *balance for the positive class*, since a violation of it would mean that people constituting positive instances in one group receive consistently lower probability estimates than people constituting positive instances in another group. In our initial criminal justice example, for instance, one of the concerns raised was that white defendants who went on to commit future crimes were assigned risk scores corresponding to lower probability estimates in aggregate; this is a violation of the condition here. There is a completely analogous property with respect to negative instances, which we could call *balance for the negative class*. These balance conditions can be viewed as generalizations of the notions that both groups should have equal false negative and false positive rates.

It is important to note that balance for the positive and negative classes, as defined here, is distinct in

crucial ways from the requirement that the average probability estimate globally over *all* members of the two groups be equal. This latter global requirement is a version of *statistical parity* [12, 4, 21, 22]. In some cases statistical parity is a central goal (and in some it is legally mandated), but the examples considered so far suggest that classification and risk assessment are much broader activities where statistical parity is often neither feasible nor desirable. Balance for the positive and negative classes, however, is a goal that can be discussed independently of statistical parity, since these two balance conditions simply ask that once we condition on the “correct” answer for a person, the chance of making a mistake on them should not depend on which group they belong to.

The present work: Trade-offs among the guarantees. Despite their different formulations, the calibration condition and the balance conditions for the positive and negative classes intuitively all seem to be asking for variants of the same general goal — that our probability estimates should have the same effectiveness regardless of group membership. One might therefore hope that it would be feasible to achieve all of them simultaneously.

Our main result, however, is that these conditions are in general incompatible with each other; they can only be simultaneously satisfied in certain highly constrained cases. Moreover, this incompatibility applies to *approximate* versions of the conditions as well.

In the remainder of this section we formulate this main result precisely, as a theorem building on a model that makes the discussion thus far more concrete.

1.1 Formulating the Goal

Let’s start with some basic definitions. As above, we have a collection of people each of whom constitutes either a positive instance or a negative instance of the classification problem. We’ll say that the *positive class* consists of the people who constitute positive instances, and the *negative class* consists of the people who constitute negative instances. For example, for criminal defendants, the positive class could consist of those defendants who will be arrested again within some fixed time window, and the negative class could consist of those who will not. The positive and negative classes thus represent the “correct” answer to the classification problem; our decision procedure does not know them, but is trying to estimate them.

Feature vectors. Each person has an associated *feature vector* σ , representing the data that we know about them. Let p_σ denote the fraction of people with feature vector σ who belong to the positive class. Conceptually, we will picture that while there is variation within the set of people who have feature vector σ , this variation is invisible to whatever decision procedure we apply; all people with feature vector σ are indistinguishable to the procedure. Our model will assume that the value p_σ for each σ is known to the procedure.¹

Groups. Each person also belongs to one of two *groups*, labeled 1 or 2, and we would like our decisions to be unbiased with respect to the members of these two groups.² In our examples, the two groups could correspond to different races or genders, or other cases where we want to look for the possibility of bias between them. The two groups have different distributions over feature vectors: a person of group t has a probability $a_{t\sigma}$ of exhibiting the feature vector σ . However, people of each group have the same probability

¹Clearly the case in which the value of p_σ is unknown is an important version of the problem as well; however, since our main results establish strong limitations on what is achievable, these limitations are only stronger because they apply even to the case of known p_σ .

²We focus on the case of two groups for simplicity of exposition, but it is straightforward to extend all of our definitions to the case of more than two groups.

p_σ of belonging to the positive class provided their feature vector is σ . In this respect, σ contains all the relevant information available to us about the person’s future behavior; once we know σ , we do not get any additional information from knowing their group as well.³

Risk Assignments. We say that an *instance* of our problem is specified by the parameters above: a feature vector and a group for each person, with a value p_σ for each feature vector, and distributions $\{a_{t\sigma}\}$ giving the frequency of the feature vectors in each group.

Informally, risk assessments are ways of dividing people up into sets based on their feature vectors σ (potentially using randomization), and then assigning each set a probability estimate that the people in this set belong to the positive class. Thus, we define a *risk assignment* to consist of a set of “bins” (the sets), where each bin is labeled with a *score* v_b that we intend to use as the probability for everyone assigned to bin b . We then create a rule for assigning people to bins based on their feature vector σ ; we allow the rule to divide people with a fixed feature vector σ across multiple bins (reflecting the possible use of randomization). Thus, the rule is specified by values $X_{\sigma b}$: a fraction $X_{\sigma b}$ of all people with feature vector σ are assigned to bin b . Note that the rule does not have access to the group t of the person being considered, only their feature vector σ . (As we will see, this does not mean that the rule is incapable of exhibiting bias between the two groups.) In summary, a risk assignment is specified by a set of bins, a score for each bin, and values $X_{\sigma b}$ that define a mapping from people with feature vectors to bins.

Fairness Properties for Risk Assignments. Within the model, we now express the three conditions discussed at the outset, each reflecting a potentially different notion of what it means for the risk assignment to be “fair.”

- (A) *Calibration within groups* requires that for each group t , and each bin b with associated score v_b , the expected number of people from group t in b who belong to the positive class should be a v_b fraction of the expected number of people from group t assigned to b .
- (B) *Balance for the negative class* requires that the average score assigned to people of group 1 who belong to the negative class should be the same as the average score assigned to people of group 2 who belong to the negative class. In other words, the assignment of scores shouldn’t be systematically more inaccurate for negative instances in one group than the other.
- (C) *Balance for the positive class* symmetrically requires that the average score assigned to people of group 1 who belong to the positive class should be the same as the average score assigned to people of group 2 who belong to the positive class.

Why Do These Conditions Correspond to Notions of Fairness?. All of these are natural conditions to impose on a risk assignment; and as indicated by the discussion above, all of them have been proposed as versions of fairness. The first one essentially asks that the scores mean what they claim to mean, even when considered separately in each group. In particular, suppose a set of scores lack the first property for some bin b , and these scores are given to a decision-maker; then if people of two different groups both belong to bin b , the decision-maker has a clear incentive to treat them differently, since the lack of calibration within groups on bin b means that these people have different aggregate probabilities of belonging to the positive class. Another way of stating the property of calibration within groups is to say that, conditioned on the bin to which an individual is assigned, the likelihood that the individual is a member of the positive class is independent of the group to which the individual belongs. This means we are justified in treating people

³As we will discuss in more detail below, the assumption that the group provides no additional information beyond σ does not restrict the generality of the model, since we can always consider instances in which people of different groups never have the same feature vector σ , and hence σ implicitly conveys perfect information about a person’s group.

with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to.

The second and third ask that if two individuals in different groups exhibit comparable future behavior (negative or positive), they should be treated comparably by the procedure. In other words, a violation of, say, the second condition would correspond to the members of the negative class in one group receiving consistently higher scores than the members of the negative class in the other group, despite the fact that the members of the negative class in the higher-scoring group have done nothing to warrant these higher scores.

We can also interpret some of the prior work around our earlier examples through the lens of these conditions. For example, in the analysis of the COMPAS risk tool for criminal defendants, the critique by Angwin et al. focused on the risk tool's violation of conditions (B) and (C); the counter-arguments established that it satisfies condition (A). While it is clearly crucial for a risk tool to satisfy (A), it may still be important to know that it violates (B) and (C). Similarly, to think in terms of the example of Internet advertising, with male and female users as the two groups, condition (A) as before requires that our estimates of ad-click probability mean the same thing in aggregate for men and women. Conditions (B) and (C) are distinct; condition (C), for example, says that a female user who genuinely wants to see a given ad should be assigned the same probability as a male user who wants to see the ad.

1.2 Determining What is Achievable: A Characterization Theorem

When can conditions (A), (B), and (C) be simultaneously achieved? We begin with two simple cases where it's possible.

- *Perfect prediction.* Suppose that for each feature vector σ , we have either $p_\sigma = 0$ or $p_\sigma = 1$. This means that we can achieve perfect prediction, since we know each person's class label (positive or negative) for certain. In this case, we can assign all feature vectors σ with $p_\sigma = 0$ to a bin b with score $v_b = 0$, and all σ with $p_\sigma = 1$ to a bin b' with score $v_{b'} = 1$. It is easy to check that all three of the conditions (A), (B), and (C) are satisfied by this risk assignment.
- *Equal base rates.* Suppose, alternately, that the two groups have the same fraction of members in the positive class; that is, the average value of p_σ is the same for the members of group 1 and group 2. (We can refer to this as the *base rate* of the group with respect to the classification problem.) In this case, we can create a single bin b with score equal to this average value of p_σ , and we can assign everyone to bin b . While this is not a particularly informative risk assignment, it is again easy to check that it satisfies fairness conditions (A), (B), and (C).

Our first main result establishes that these are in fact the only two cases in which a risk assignment can achieve all three fairness guarantees simultaneously.

Theorem 1.1 *Consider an instance of the problem in which there is a risk assignment satisfying fairness conditions (A), (B), and (C). Then the instance must either allow for perfect prediction (with p_σ equal to 0 or 1 for all σ) or have equal base rates.*

Thus, in every instance that is more complex than the two cases noted above, there will be some natural fairness condition that is violated by any risk assignment. Moreover, note that this result applies regardless of how the risk assignment is computed; since our framework considers risk assignments to be arbitrary functions from feature vectors to bins labeled with probability estimates, it applies independently of the method — algorithmic or otherwise — that is used to construct the risk assignment.

The conclusions of the first theorem can be relaxed in a continuous fashion when the fairness conditions are only approximate. In particular, for any $\varepsilon > 0$ we can define ε -approximate versions of each of conditions (A), (B), and (C) (specified precisely in the next section), each of which requires that the corresponding equalities between groups hold only to within an error of ε . For any $\delta > 0$, we can also define a δ -approximate version of the equal base rates condition (requiring that the base rates of the two groups be within an additive δ of each other) and a δ -approximate version of the perfect prediction condition (requiring that in each group, the average of the expected scores assigned to members of the positive class is at least $1 - \delta$; by the calibration condition, this can be shown to imply a complementary bound on the average of the expected scores assigned to members of the negative class).

In these terms, our approximate version of Theorem 1.1 is the following.

Theorem 1.2 *There is a continuous function f , with $f(x)$ going to 0 as x goes to 0, so that the following holds. For all $\varepsilon > 0$, and any instance of the problem with a risk assignment satisfying the ε -approximate versions of fairness conditions (A), (B), and (C), the instance must satisfy either the $f(\varepsilon)$ -approximate version of perfect prediction or the $f(\varepsilon)$ -approximate version of equal base rates.*

Thus, anything that approximately satisfies the fairness constraints must approximately look like one of the two simple cases identified above.

Finally, in connection to Theorem 1.1, we note that when the two groups have equal base rates, then one can ask for the most accurate risk assignment that satisfies all three fairness conditions (A), (B), and (C) simultaneously. Since the risk assignment that gives the same score to everyone satisfies the three conditions, we know that at least one such risk assignment exists; hence, it is natural to seek to optimize over the set of all such assignments. We consider this algorithmic question in the final technical section of the paper.

To reflect a bit further on our main theorems and what they suggest, we note that our intention in the present work isn't to make a recommendation on how conflicts between different definitions of fairness should be handled. Nor is our intention to analyze which definitions of fairness are violated in particular applications or datasets. Rather, our point is to establish certain unavoidable trade-offs between the definitions, regardless of the specific context and regardless of the method used to compute risk scores. Since each of the definitions reflect (and have been proposed as) natural notions of what it should mean for a risk score to be fair, these trade-offs suggest a striking implication: that outside of narrowly delineated cases, any assignment of risk scores can in principle be subject to natural criticisms on the grounds of bias. This is equally true whether the risk score is determined by an algorithm or by a system of human decision-makers.

Special Cases of the Model. Our main results, which place strong restrictions on when the three fairness conditions can be simultaneously satisfied, have more power when the underlying model of the input is more general, since it means that the restrictions implied by the theorems apply in greater generality. However, it is also useful to note certain special cases of our model, obtained by limiting the flexibility of certain parameters in intuitive ways. The point is that our results apply *a fortiori* to these more limited special cases.

First, we have already observed one natural special case of our model: cases in which, for each feature vector σ , only members of one group (but not the other) can exhibit σ . This means that σ contains perfect information about group membership, and so it corresponds to instances in which risk assignments would have the potential to use knowledge of an individual's group membership. Note that we can convert any instance of our problem into a new instance that belongs to this special case as follows. For each feature vector σ , we create two new feature vectors $\sigma^{(1)}$ and $\sigma^{(2)}$; then, for each member of group 1 who had feature vector σ , we assign them $\sigma^{(1)}$, and for each member of group 2 who had feature vector σ , we assign them

$\sigma^{(2)}$. The resulting instance has the property that each feature vector is associated with members of only one group, but it preserves the essential aspects of the original instance in other respects.

Second, we allow risk assignments in our model to split people with a given feature vector σ over several bins. Our results also therefore apply to the natural special case of the model with *integral* risk assignments, in which all people with a given feature σ must go to the same bin.

Third, our model is a generalization of binary classification, which only allows for 2 bins. Note that although binary classification does not explicitly assign scores, we can consider the probability that an individual belongs to the positive class given that they were assigned to a specific bin to be the score for that bin. Thus, our results hold in the traditional binary classification setting as well.

Data-Generating Processes. Finally, there is the question of where the data in an instance of our problem comes from. Our results do not assume any particular process for generating the positive/negative class labels, feature vectors, and group memberships; we simply assume that we are given such a collection of values (regardless of where they came from), and then our results address the existence or non-existence of certain risk assignments for these values.

This increases the generality of our results, since it means that they apply to any process that produces data of the form described by our model. To give an example of a natural generative model that would produce instances with the structure that we need, one could assume that each individual starts with a “hidden” class label (positive or negative), and a feature vector σ is then probabilistically generated for this individual from a distribution that can depend on their class label and their group membership. (If feature vectors produced for the two groups are disjoint from one another, then the requirement that the value of p_σ is independent of group membership given σ necessarily holds.) Since a process with this structure produces instances from our model, our results apply to data that arises from such a generative process.

It is also interesting to note that the basic set-up of our model, with the population divided across a set of feature vectors for which race provides no additional information, is in fact a very close match to the information one gets from the output of a well-calibrated risk tool. In this sense, one setting for our model would be the problem of applying post-processing to the output of such a risk tool to ensure additional fairness guarantees. Indeed, since much of the recent controversy about fair risk scores has involved risk tools that are well-calibrated but lack the other fairness conditions we consider, such an interpretation of the model could be a useful way to think about how one might work with these tools in the context of a broader system.

1.3 Further Related Work

Mounting concern over discrimination in machine learning has led to a large body of new work seeking to better understand and prevent it. Barcas and Selbst survey a range of ways in which data-analysis algorithms can lead to discriminatory outcomes [3], and review articles by Romei and Ruggieri [25] and Zliobaite [30] survey data-analytic and algorithmic methods for measuring discrimination.

Kamiran and Calders [21] and Hajian and Domingo-Ferrer [18] seek to modify datasets to remove any information that might permit discrimination. Similarly, Zemel et al. look to learn fair intermediate representations of data while preserving information needed for classification [29]. Joseph et al. consider how fairness issues can arise during the process of learning, modeling this using a multi-armed bandit framework [20].

Fair prediction with disparate impact: A study of bias in recidivism prediction instruments

Alexandra Chouldechova ^{*}

Last revised: February 8, 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing^{1,2,3}. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If the use of RPI's is to become commonplace, it is especially important to ensure that the instruments are free from discriminatory biases that could result in unethical practices and inequitable outcomes for different groups.

In a recent widely popularized investigation conducted by a team at ProPublica, Angwin et al.⁴ studied an RPI called COMPAS^a, concluding that it is biased against black defendants. The authors

^{*}Heinz College, Carnegie Mellon University

^aCOMPAS⁵ is a risk assessment instrument developed by Northpointe Inc.. Of the 22 scales that COMPAS provides, the Recidivism risk and Violent Recidivism risk scales are the most widely used. The empirical results in this paper are based on decile scores coming from the COMPAS Recidivism risk scale.

found that the likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of white defendants. Similarly, the likelihood of a recidivating black defendant being assessed as low risk is nearly half that of white defendants. In technical terms, these findings indicate that the COMPAS instrument has considerably higher false positive rates and lower false negative rates for black defendants than for white defendants.

ProPublica’s analysis has met with much criticism from both the academic community and from the Northpointe corporation. Much of the criticism has focussed on the particular choice of fairness criteria selected for the investigation. Flores et al.⁶ argue that the correct approach for assessing RPI bias is instead to check for *calibration*, a fairness criterion that they show COMPAS satisfies. Northpointe in their response⁷ argue for a still different approach that checks for a fairness criterion termed *predictive parity*, which they demonstrate COMPAS also satisfies. We provide precise definitions and a more in-depth discussion of these and other fairness criteria in Section 2.1.

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al.⁴ are a direct consequence of applying an RPI that that satisfies predictive parity to a population in which recidivism prevalence^a differs across groups. Our main contribution is twofold. (1) First, we make precise the connection between the predictive parity criterion and error rates in classification. (2) Next, we demonstrate how using an RPI that has different false positive and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties. Throughout our discussion we use the term *disparate impact* to refer to settings where a penalty policy has unintended disproportionate adverse impact on a particular group.

It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. A risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used. In this paper we consider hypothetical use cases in which we are able to directly connect particular fairness properties of an RPI to a measure of disparate impact. We present both theoretical and empirical results to illustrate how disparate impact can arise.

1.1 Outline of paper

We begin in Section 2 by providing some background on several of the different fairness criteria that have appeared in recent literature. We then proceed to demonstrate that an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups. In Section 3 we analyse a simple risk assessment-based sentencing policy and show how differences in false positive and false negative rates can result in disparate impact under this policy. In Section 3.3 we back up our theoretical analysis by presenting some empirical results based on the data made available by the ProPublica investigators. We conclude with a discussion of the issues that biased data presents for the arguments put forth in this paper.

^a*Prevalence*, also termed the *base rate*, is the proportion of individuals who recidivate in a given population.

1.2 Data description and setup

The empirical results in this paper are based on the Broward County data made publicly available by ProPublica⁸. This data set contains COMPAS recidivism risk decile scores, 2-year recidivism outcomes, and a number of demographic and crime-related variables on individuals who were scored in 2013 and 2014. We restrict our attention to the subset of defendants whose race is recorded as African-American (b) or Caucasian (w).^a After applying the same data pre-processing and filtering as reported in the ProPublica analysis, we are left with a data set on $n = 6150$ individuals, of whom $n_b = 3696$ are African-American and $n_c = 2454$ are Caucasian.

2 Assessing fairness

2.1 Background

We begin by with some notation. Let $S = S(x)$ denote the risk score based on covariates $X = x \in \mathbb{R}^p$, with higher values of S corresponding to higher levels of assessed risk. We will interchangeably refer to S as a *score* or an *instrument*. For simplicity, our discussion of fairness criteria will focus on a setting where there exist just two groups. We let $R \in \{b, w\}$ denote the group to which an individual belongs, and do not preclude R from being one of the elements of X . We denote the outcome indicator by $Y \in \{0, 1\}$, with $Y = 1$ indicating that the given individual goes on to recidivate. Lastly, we introduce the quantity s_{HR} , which denotes the high-risk score threshold. Defendants whose score S exceeds s_{HR} will be referred to as *high-risk*, while the remaining defendants will be referred to as *low-risk*.

With this notation in hand, we now proceed to define and discuss several fairness criteria that commonly appear in the literature, beginning with those mentioned in the introduction. We indicate cases where a given criterion is known to us to also commonly appear under some other name. All of the criteria presented below can also be assessed *conditionally* by further conditioning on some covariates in X . We discuss this point in greater detail in Section 3.1.

Definition 1 (Calibration). A score $S = S(x)$ is said to be *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. That is, if for all values of s ,

$$\mathbb{P}(Y = 1 | S = s, R = b) = \mathbb{P}(Y = 1 | S = s, R = w). \quad (2.1)$$

Within the educational and psychological testing and assessment literature, the notion of *calibration* features among the widely accepted and adopted standards for empirical fairness assessment. In this literature, an instrument that is *well-calibrated* is referred to as being *free from predictive bias*. This criterion has recently been applied to the PCRA^b instrument, with initial findings suggesting that calibration is satisfied with respect race^{10,11}, but not with respect to gender¹². In

^aThere are 6 racial groups represented in the data. 85% of individuals are either African-American or Caucasian.

^bThe Post Conviction Risk Assessment (PCRA) tool was developed by the Administrative Office of the United States Courts for the purpose of improving “the effectiveness and efficiency of post-conviction supervision”⁹

their response to the ProPublica investigation, Flores et al.⁶ verify that COMPAS is well-calibrated using logistic regression modeling.

Definition 2 (Predictive parity). A score $S = S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if,

$$\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = w). \quad (2.2)$$

Predictive parity at a given threshold s_{HR} amounts to requiring that the *positive predictive value* (PPV) of the classifier $\hat{Y} = \mathbf{1}_{S > s_{\text{HR}}}$ be the same across groups. While predictive parity and calibration look like very similar criteria, well-calibrated scores can fail to satisfy predictive parity at a given threshold. This is because the relationship between (2.2) and (2.1) depends on the conditional distribution of $S \mid R = r$, which can differ across groups in ways that result in PPV imbalance. In the simple case where S itself is binary, a score that is well-calibrated will also satisfy predictive parity. Northpointe's refutation⁷ of the ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choices of interest.

Definition 3 (Error rate balance). A score $S = S(x)$ satisfies *error rate balance* at a threshold s_{HR} if the false positive and false negative error rates are equal across groups. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = w), \quad \text{and} \quad (2.3)$$

$$\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = w), \quad (2.4)$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates.

ProPublica's analysis considered a threshold of $s_{\text{HR}} = 4$, which they showed leads to considerable imbalance in both false positive and false negative rates. While this choice of cutoff met with some criticism, we will see later in this section that error rate imbalance persists—indeed, must persist—for any choice of cutoff at which the score satisfies the predictive parity criterion. Error rate balance is also closely connected to the notions of *equalized odds* and *equal opportunity* as introduced in the recent work of Hardt et al.¹³.

Definition 4 (Statistical parity). A score $S = S(x)$ satisfies *statistical parity* at a threshold s_{HR} if the proportion of individuals classified as high-risk is the same for each group. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid R = b) = \mathbb{P}(S > s_{\text{HR}} \mid R = w) \quad (2.5)$$

Statistical parity also goes by the name of *equal acceptance rates*¹⁴ or *group fairness*¹⁵, though it should be noted that these terms are in many cases not used synonymously. While our discussion focusses primarily on first three fairness criteria, statistical parity is widely used within the machine learning community and may be the criterion with which many readers are most familiar^{16,17}. Statistical parity is well-suited to contexts such as employment or admissions, where it may be desirable or required by law or regulation to employ or admit individuals in equal proportion across racial, gender, or geographical groups. It is, however, a difficult criterion to motivate in the recidivism prediction setting, and thus will not be further considered in this work.

2.2 Further related work

Though the study of discrimination in decision making and predictive modeling is rapidly evolving, it also has a long and rich multidisciplinary history. Romei and Ruggieri¹⁸ provide an excellent overview of some of the work in this broad subject area. The recent work of Barocas and Selbst¹⁹ offers a broad examination of algorithmic fairness framed within the context of anti-discrimination laws governing employment practices. Hannah-Moffat²⁰, Skeem²¹, and Monahan and Skeem²² examine legal and ethical issues relating specifically to the use of risk assessment instruments in sentencing, citing the potential for race and gender discrimination as a major concern.

In work concurrent with our own, several other researchers have also investigated the compatibility of different notions of fairness. Kleinberg et al.²³ show that calibration cannot be satisfied simultaneously with the fairness criteria of *balance for the negative class* and *balance for the positive class*. Translated into the present context, the latter criteria require that the average score assigned to non-recidivists (the negative class) should be the same for both groups, and that the same should hold among recidivists (the positive class). The work of Corbett-Davies et al.²⁴ closely parallels the results that we present in Section 2.3, reaching the same conclusion regarding the incompatibility of predictive parity and error rate balance in the setting of unequal prevalence.

2.3 Predictive parity, false positive rates, and false negative rates

In this section we present our first main result, which establishes that predictive parity is incompatible with error rate balance when prevalence differs across groups. To better motivate the discussion, we begin by presenting an empirical fairness assessment of the COMPAS RPI. Figure 1 shows plots of the observed recidivism rates and error rates corresponding to the fairness notions of calibration, predictive parity, and error rate balance. We see that the COMPAS RPI is (approximately) well-calibrated, and also satisfies predictive parity provided that the high-risk cutoff s_{HR} is 4 or greater. However, COMPAS fails on both false positive and false negative error rate balance across the range of high-risk cutoffs.

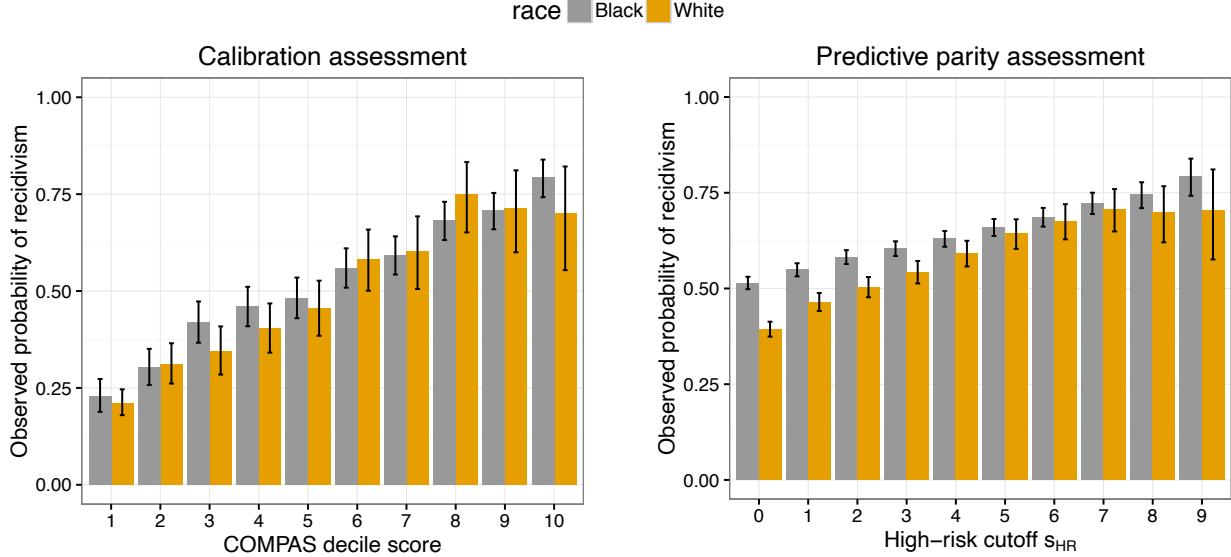
Angwin et al.⁴ focussed on a high-risk cutoff of $s_{HR} = 4$ for their analysis, which some critics have argued is too low, suggesting that $s_{HR} = 7$ is more suitable. As can be seen from Figures 1c and 1d, significant error rate imbalance persists at this cut-off as well. Moreover, the error rates achieved at so high a cutoff are at odds with evidence suggesting that the use of RPI's is of interest in settings where false negatives have a higher cost than false positives, with relative cost estimates ranging from 2.6 to upwards of 15.^{25,26}

As we now proceed to show, the error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context. When the recidivism prevalence—i.e., the base rate $\mathbb{P}(Y = 1 | R = r)$ —differs across groups, any instrument that satisfies predictive parity at a given threshold s_{HR} *must* have imbalanced false positive or false negative errors rates at that threshold. To understand why predictive parity and error rate balance are mutually exclusive in the setting of unequal recidivism prevalence, it is instructive to think of how these quantities are all related.

Given a particular choice of s_{HR} , we can summarize an instrument's performance in terms of a confusion matrix, as shown in Table 1 below.

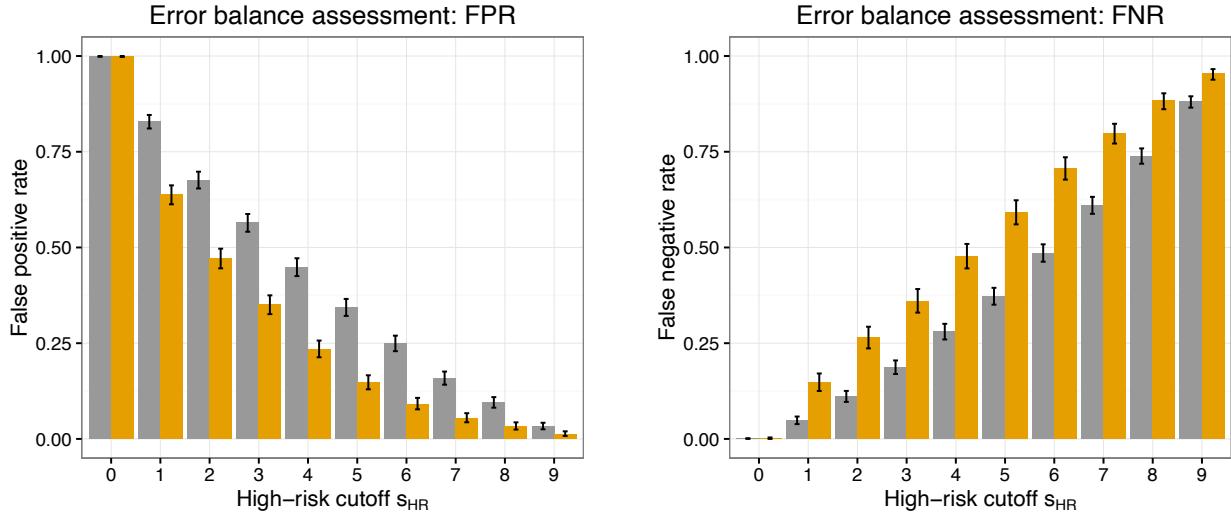
All of the fairness metrics presented in Section 2.1 can be thought of as imposing constraints on

the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to



(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \dots, 10\}$.

(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{HR}, R = r)$ for values of the high-risk cutoff $s_{HR} \in \{0, \dots, 9\}$



(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{HR} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{HR} \in \{0, \dots, 9\}$

(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \leq s_{HR} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{HR} \in \{0, \dots, 9\}$

Figure 1: Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

	Low-Risk	High-Risk
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Table 1: T/F denote True/False and N/P denote Negative/Positive. For instance, FP is the number of false positives: individuals who are classified as high-risk but who do not reoffend.

show that the prevalence (p), positive predictive value (PPV), and false positive and negative error rates (FPR, FNR) are related via the equation

$$\text{FPR} = \frac{p}{1-p} \frac{1 - \text{PPV}}{\text{PPV}} (1 - \text{FNR}). \quad (2.6)$$

From this simple expression we can see that if an instrument satisfies predictive parity—that is, if the PPV is the same across groups—but the prevalence differs between groups, the instrument cannot achieve equal false positive and false negative rates across those groups.

This observation enables us to better understand why we observe such large discrepancies in FPR and FNR between black and white defendants in Figure 1. The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants. Thus at any threshold s_{HR} where the COMPAS RPI satisfies predictive parity, equation (2.6) tells us that some level of imbalance in the error rates must exist. Since not all of the fairness criteria can be satisfied at the same time, it becomes important to understand the potential impact of failing to satisfy particular criteria. This question is explored in the context of a hypothetical risk-based sentencing framework in the next section.

3 Assessing impact

In this section we show how differences in false positive and false negative rates can result in disparate impact under policies where a high-risk assessment results in a stricter penalty for the defendant. Such situations may arise when risk assessments are used to inform bail, parole, or sentencing decisions. In Pennsylvania and Virginia, for instance, statutes permit the use of RPI’s in sentencing, provided that the sentence ultimately falls within accepted guidelines¹. We use the term “penalty” somewhat loosely in this discussion to refer to outcomes both in the pre-trial and post-conviction phase of legal proceedings. For instance, even though pre-trial outcomes such as the amount at which bail is set are not punitive in a legal sense, we nevertheless refer to bail amount as a “penalty” for the purpose of our discussion.

There are notable cases where RPI’s are used for the express purpose of informing risk reduction efforts. In such settings, individuals assessed as high risk receive what may be viewed as a benefit rather than a penalty. The PCRA score, for instance, is intended to support precisely this type of decision-making at the federal courts level¹¹. Our analysis in this section specifically addresses use cases where high-risk individuals receive stricter penalties.

To begin, consider a setting in which guidelines indicate that a defendant is to receive a penalty

RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer^{1,2*}, Brian Powers³, Christine Vogeli⁴, Sendhil Mullainathan^{5*†}

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remediating this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (1–3). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (4), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (5), and image searches for professions such as CEO produce fewer images of women (6). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (7, 8), and natural language processing algorithms encode language in gendered ways (9).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work “from the outside,” often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with

researcher-created algorithms (10–13). Without an algorithm’s training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for “high-risk care management” programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (14–17). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (18, 19).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (20). Developers then build algorithms

that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm’s predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (21).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

¹School of Public Health, University of California, Berkeley, Berkeley, CA, USA. ²Department of Emergency Medicine, Brigham and Women’s Hospital, Boston, MA, USA.

³Department of Medicine, Brigham and Women’s Hospital, Boston, MA, USA. ⁴Morgan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. ⁵Booth School of Business, University of Chicago, Chicago, IL, USA.

*These authors contributed equally to this work.

†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

and ethnic identities. Our main sample thus consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively (1 patient-year represents data collected for an individual patient in a calendar year). The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female (Table 1).

For these patients, we obtained algorithmic risk scores generated for each patient-year. In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program. Patients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.

Many existing metrics of algorithmic bias may apply to this scenario. Some definitions focus on calibration [i.e., whether the realized value of some variable of interest Y matches the risk score R (2, 22, 23)]; others on statistical parity of some decision D influenced by the algorithm (10); and still others on balance of average predictions, conditional on the realized outcome (22). Given this multiplicity and the growing recognition that not all conditions can be simultaneously satisfied (3, 10, 22), we focus on metrics most relevant to the real-world use of the algorithm, which are related to calibration bias [formally, comparing Blacks B and Whites W , $E[Y|R, W] = E[Y|R, B]$ indicates the absence of bias (here, E is the expectation operator)]. The algorithm's stated goal is to predict complex health needs for the purpose of targeting an intervention that manages those needs. Thus, we compare the algorithmic risk score for patient i in year t ($R_{i,t}$), formed on the basis of claims data $X_{i,(t-1)}$ from the prior year, to data on patients' realized health $H_{i,t}$, assessing how well the algorithmic risk score is calibrated across race for health outcomes $H_{i,t}$. We also ask how well the algorithm is calibrated for costs $C_{i,t}$.

To measure H , we link predictions to a wide range of outcomes in electronic health record data, including all diagnoses (in the form of International Classification of Diseases codes) as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses. To measure C , we link predictions to insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs. These data, and the rationale for the specific measures of H used in this study, are described in more detail in the supplementary materials.

Health disparities conditional on risk score

We begin by calculating an overall measure of health status, the number of active chronic conditions [or "comorbidity score," a metric used extensively in medical research (24) to provide a comprehensive view of a patient's health (25)] by race, conditional on algorithmic risk score. Fig. 1A shows that, at the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites. We can quantify these differences by choosing one point on the x axis that corresponds to

a very-high-risk group (e.g., patients at the 97th percentile of risk score, at which patients are auto-identified for program enrollment), where Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$).

What do these prediction differences mean for patients? Algorithm scores are a key input to decisions about future enrollment in a care coordination program. So as we might expect, with less-healthy Blacks scored at similar risk scores to more-healthy Whites, we find evidence

Table 1. Descriptive statistics on our sample, by race. BP, blood pressure; LDL, low-density lipoprotein.

	White	Black
<i>n</i> (patient-years)	88,080	11,929
<i>n</i> (patients)	43,539	6079
<i>Demographics</i>		
Age	51.3	48.6
Female (%)	62	69
<i>Care management program</i>		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
<i>Care utilization</i>		
Actual cost	\$7540	\$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
<i>Mean biomarker values</i>		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
<i>Active chronic illnesses (comorbidities)</i>		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08
Hypothyroid	0.09	0.05
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Cancer	0.07	0.06
Depression	0.06	0.08
Anemia	0.05	0.10
Arthritis	0.04	0.04
Renal failure	0.03	0.07
Electrolyte disorder	0.03	0.05
Heart failure	0.03	0.05
Psychosis	0.03	0.05
Valvular disease	0.03	0.02
Stroke	0.02	0.03
Peripheral vascular disease	0.02	0.02
Diabetes, complicated	0.02	0.07
Heart attack	0.01	0.02
Liver disease	0.01	0.02

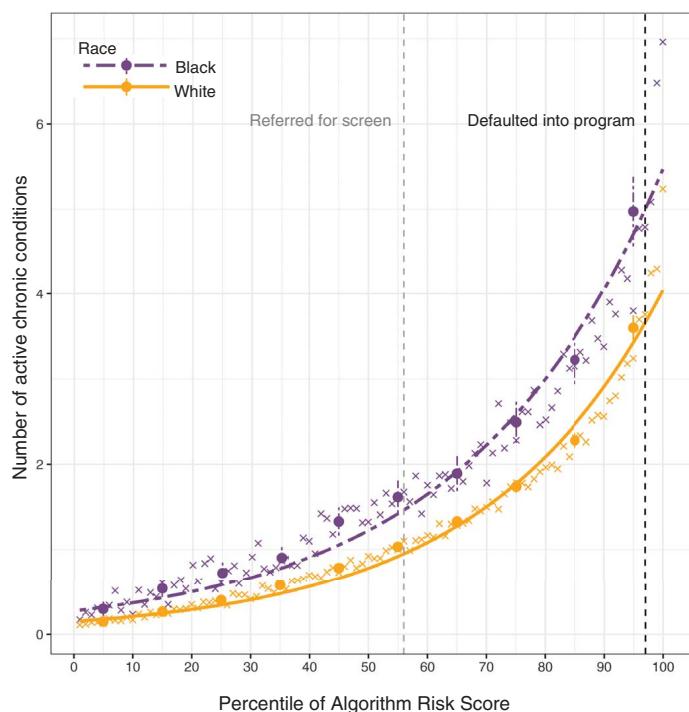
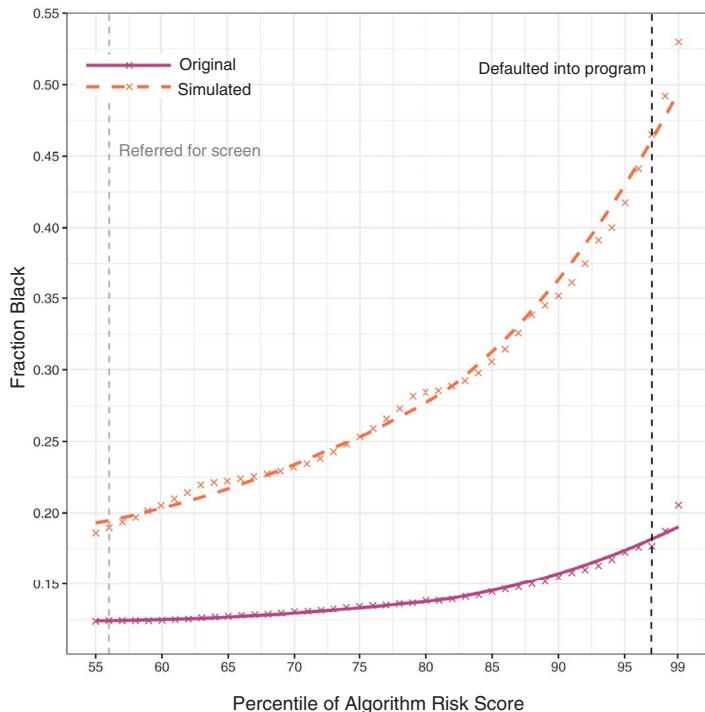
A**B**

Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (A) Mean number of chronic conditions by race, plotted against algorithm risk score. (B) Fraction of Black patients at or above a given risk score for the original algorithm (“original”) and for a simulated scenario that removes algorithmic bias (“simulated”): at each threshold of risk, defined at a given percentile on the x axis, healthier Whites above the threshold are

of substantial disparities in program screening. We quantify this by simulating a counterfactual world with no gap in health conditional on risk. Specifically, at some risk threshold α , we identify the supramarginal White patient (i) with $R_i > \alpha$ and compare this patient’s health to that of the inframarginal Black patient (j) with $R_j < \alpha$. If $H_i > H_j$, as measured by number of chronic medical conditions, we replace the (healthier, but supramarginal) White patient with the (sicker, but inframarginal) Black patient. We repeat this procedure until $H_i = H_j$, to simulate an algorithm with no predictive gap between Blacks and Whites. Fig. 1B shows the results: At all risk thresholds α above the 50th percentile, this procedure would increase the fraction of Black patients. For example, at $\alpha = 97$ th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.

We then turn to a more multidimensional picture of the complexity and severity of patients’ health status, as measured by biomarkers that index the severity of the most common chronic illnesses in our sample (as shown in Table 1). This allows us to identify patients who might derive a great deal of benefit from care management programs—e.g., patients with severe

diabetes who are at risk of catastrophic complications if they do not lower their blood sugar (18, 26). (The materials and methods section describes several experiments to rule out a large effect of the program on these health measures in year t ; had there been such an effect, we could not easily use the measures to assess the accuracy of the algorithm’s predictions on health, because the program is allocated as a function of algorithm score.) Across all of these important markers of health needs—severity of diabetes, high blood pressure, renal failure, cholesterol, and anemia—we find that Blacks are substantially less healthy than Whites at any level of algorithm predictions, as shown in Fig. 2. Blacks have more-severe hypertension, diabetes, renal failure, and anemia, and higher cholesterol. The magnitudes of these differences are large: For example, differences in severity of hypertension (systolic pressure: 5.7 mmHg) and diabetes [glycated hemoglobin (HbA1c): 0.6%] imply differences in all-cause mortality of 7.6% (27) and 30% (28), respectively, calculated using data from clinical trials and longitudinal studies.

Mechanism of bias

An unusual aspect of our dataset is that we observe the algorithm’s inputs and outputs

replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The \times symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

as well as its objective function, providing us a unique window into the mechanisms by which bias arises. In our setting, the algorithm takes in a large set of raw insurance claims data $X_{i,t-1}$ (features) over the year $t - 1$: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. Notably, the algorithm specifically excludes race.

The algorithm uses these data to predict $Y_{i,t}$ (i.e., the label). In this instance, the algorithm takes total medical expenditures (for simplicity, we denote “costs” C_t) in year t as the label. Thus, the algorithm’s prediction on health needs is, in fact, a prediction on health costs.

As a first check on this potential mechanism of bias, we calculate the distribution of realized costs C versus predicted costs R . By this metric, one could call the algorithm unbiased. Fig. 3A shows that, at every level of algorithm-predicted risk, Blacks and Whites have (roughly) the same costs the following year. In other words, the algorithm’s predictions are well calibrated across races. For example, at the median risk score, Black patients had costs of \$5147 versus \$4995 for Whites (U.S. dollars); in the top 5% of algorithm-predicted risk, costs were \$35,541 for Blacks versus \$34,059 for Whites.

Because these programs are used to target patients with high costs, these results are largely inconsistent with algorithmic bias, as measured by calibration: Conditional on risk score, predictions do not favor Whites or Blacks anywhere in the risk distribution.

To summarize, we find substantial disparities in health conditional on risk but little disparity in costs. On the one hand, this is surprising: Health care costs and health needs are highly correlated, as sicker patients need and receive more care, on average. On the other hand, there are many opportunities for a wedge to creep in between needing health care and receiving health care—and crucially, we find that wedge to be correlated with race, as shown in Fig. 3B. At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than Whites—on average, \$1801 less per year, holding constant the number of chronic illnesses (or \$1144 less, if we instead hold constant the specific individual illnesses that contribute to the sum). Table S2 also shows that Black patients generate very different kinds of costs: for example, fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis. These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities. As a result, accurate prediction of costs necessarily means being racially biased on health.

How might these disparities in cost arise? The literature broadly suggests two main potential channels. First, poor patients face substantial barriers to accessing health care, even when enrolled in insurance plans. Although the population we study is entirely insured, there are many other mechanisms by which poverty can lead to disparities in use of health care: geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care (29–31). To the extent that race and socioeconomic status are correlated, these factors will differentially affect Black patients. Second, race could affect costs directly via several channels: direct (“taste-based”) discrimination, changes to the doctor–patient relationship, or others. A recent trial randomly assigned Black patients to a Black or White primary care provider and found significantly higher uptake of recommended preventive care when the provider was Black (32). This is perhaps the most rigorous demonstration of this effect, and it fits with a larger literature on potential mechanisms by which race can affect health care directly. For example, it has long been documented that Black patients have reduced trust in the health care system (33), a fact that some studies trace to the revelations of the Tuskegee study and other adverse experiences (34). A substantial

literature in psychology has documented physicians’ differential perceptions of Black patients, in terms of intelligence, affiliation (35), or pain tolerance (36). Thus, whether it is communication, trust, or bias, something about the interactions of Black patients with the health care system itself leads to reduced use of health care. The collective effect of these many channels is to lower health spending substantially for Black

patients, conditional on need—a finding that has been appreciated for at least two decades (37).

Problem formulation

Our findings highlight the importance of the choice of the label on which the algorithm is trained. On the one hand, the algorithm manufacturer’s choice to predict future costs is reasonable: The program’s goal, at least in part, is

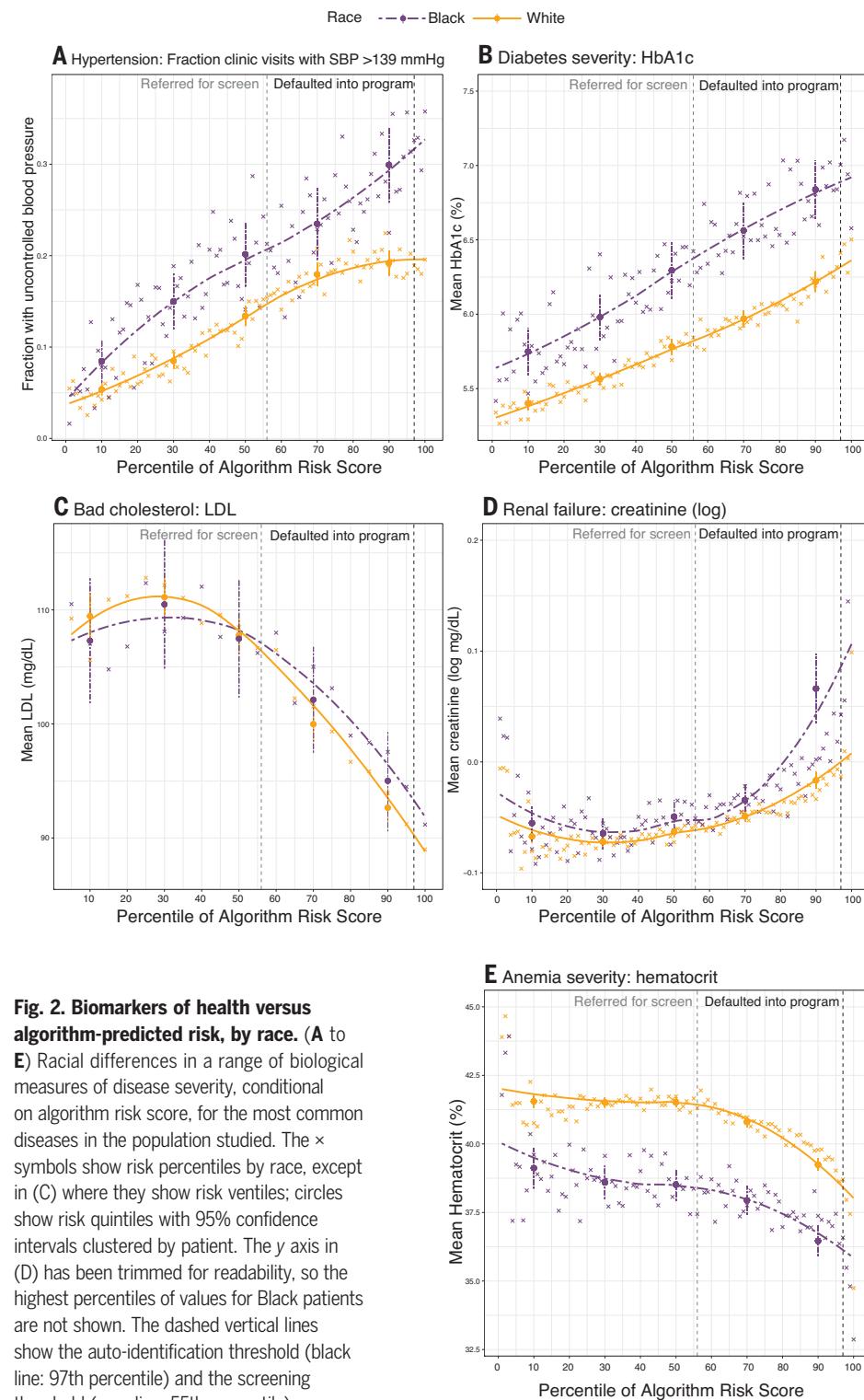


Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (A to E) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The \times symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The y axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

to reduce costs, and it stands to reason that patients with the greatest future costs could have the greatest benefit from the program. As noted in the supplementary materials, the manufacturer is not alone. Although the details of individual algorithms vary, the cost label reflects the industry-wide approach. For example, the Society of Actuaries's comprehensive evaluation of the 10 most widely used algorithms, including the particular algorithm we study, used cost prediction as its accuracy metric (21). As noted in the report, the enthusiasm for cost prediction is not restricted to industry: Similar algorithms are developed and used by non-profit hospitals, academic groups, and governmental agencies, and are often described in academic literature on targeting population health interventions (18, 19).

On the other hand, future cost is by no means the only reasonable choice. For example, the evidence on care management programs shows that they do not operate to reduce costs globally. Rather, these programs primarily work to prevent acute health decompensations that lead to catastrophic health care utilization (indeed, they actually work to increase other categories of costs, such as primary care and home health assistance; see table S2). Thus avoidable future costs, i.e., those related to emergency visits and hospitalizations, could be a useful label to predict. Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions. Because the program ultimately operates to improve the management of these conditions, patients with the most encounters related to them could also be a promising group on which to deploy preventative interventions.

The dilemma of which label to choose relates to a growing literature on “problem formulation” in data science: the task of turning an often amorphous concept we wish to predict into a concrete variable that can be predicted in a given dataset (38). Problems in health seem particularly challenging: Health is, by nature, holistic and multidimensional, and there is no single, precise way to measure it. Health care costs, though well measured and readily available in insurance claims data, are also the result of a complex aggregation process with a number of distortions due to structural inequality, incentives, and inefficiency. So although the choice of label is perhaps the single most important decision made in the development of a prediction algorithm, in our setting and in many others, there is often a confusingly large array of different options, each with its own profile of costs and benefits.

Experiments on label choice

Through a series of experiments with our dataset, we can gain some insight into how label choice affects both predictive performance and racial bias. We develop three new predictive algorithms, all trained in the same way, to predict the following outcomes: total cost in year t (this tailors cost predictions to our own dataset rather than the national training set), avoidable cost in year t (due to emergency visits and hospitalizations), and health in year t (measured by the number of chronic conditions that flare up in that year). We train all models in a random $\frac{2}{3}$ training set and show all results only from the $\frac{1}{3}$ holdout set. Furthermore, as with the original algorithm, we exclude race from the feature set (more details are in the materials and methods).

Table 2 shows the results of these experiments. The first finding is that all algorithms perform reasonably well for predicting not only the outcome on which they were trained but also the other outcomes: The concentration of realized outcomes in those at or above the 97th percentile is notably similar for all algorithms across all outcomes. The largest difference in performance across algorithms is seen for cost prediction: Of all costs in the holdout set, the fraction generated by those at or above the 97th percentile is 16.5% for the cost predictor versus 12.1% for the predictor

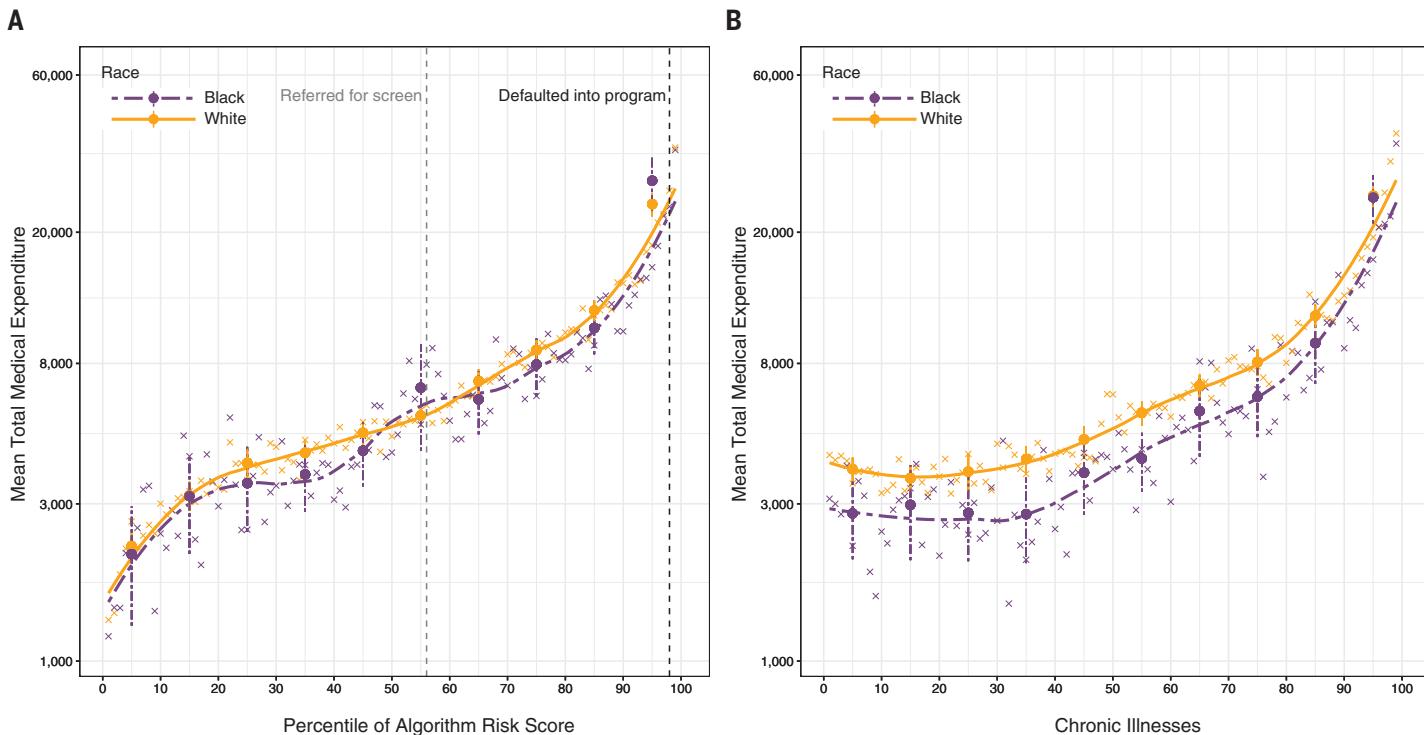


Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (A) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (B) Total medical expenditures by race, conditional on number of chronic conditions. The \times symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The y axis uses a log scale.

of chronic conditions. We then test for label choice bias, defined analogously to calibration bias above: For two algorithms trained to predict Y and Y' , and using a threshold τ indexing a (similarly sized) high-risk group, we would test $p[B|R > \tau] = p[B|R' > \tau]$ (here, p denotes probability and B represents Black patients).

We find that the racial composition of this highest-risk group varies far more across algorithms: The fraction of Black patients at or above these risk levels ranges from 14.1% for the cost predictor to 26.7% for the predictor of chronic conditions. Thus, although there could be many reasonable choices of label—all predictions are highly correlated, and any could be justified as a measure of patients’ likely benefit from the program—they have markedly different implications in terms of bias, with nearly twofold variation in composition of Black patients in the highest-risk groups.

Relation to human judgment

As noted above, the algorithm is not used for program enrollment decisions in isolation. Rather, it is used as a screening tool, in part to alert primary care doctors to high-risk

patients. Specifically, for patients at or above a certain level of predicted risk (the 55th percentile), doctors are presented with contextual information from patients’ electronic health records and insurance claims and are prompted to consider enrolling them in the program. Thus, realized enrollment decisions largely reflect how doctors respond to algorithmic predictions, along with other administrative factors related to eligibility (for instance, primary care practice site, residence outside of a nursing home, and continual enrollment in an insurance plan).

Table 3 shows statistics on those enrolled in the program, accounting for 1.3% of observations in our sample: The enrolled individuals are 19.2% Black (versus 11.9% Black in our entire sample) and account for 2.9% of all costs and 3.3% of all active chronic conditions in the population as a whole. We then perform four counterfactual simulations to put these numbers in context; naturally, these simulations use only observable factors, not the many unobserved administrative and human factors that also affect enrollment. First, we calculate the realized program enrollment rate within each percentile of the original algorithm’s pre-

dicted risk bins and randomly sample patients in each bin for enrollment. This simulation, which mimics “race-blind” enrollment conditional on algorithm score, would yield an enrolled population that is 18.3% Black (versus 19.2% observed; $P = 0.8348$). Second, rather than randomly sampling, we sample those with the highest predicted number of active chronic conditions within a risk bin (using our experimental algorithm described above); this would yield a population that is 26.9% Black. Finally, we compare this to simply assigning those with the highest predicted costs, or the highest number of active chronic conditions, to the program (also using our own algorithms detailed above), which would yield 17.2 and 29.2% Black patients, respectively. Thus, although doctors do redress a small part of the algorithm’s bias, they do so far less than an algorithm trained on a different label.

Discussion

Bias attributable to label choice—the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label—is a useful framework through which to understand bias in algorithms, both

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)			Fraction of Black patients in group with highest risk (SE)	
	Total costs	Avoidable costs	Active chronic conditions		
Total costs	0.165 (0.003)	0.187 (0.003)	0.105 (0.002)	0.141 (0.003)	
Avoidable costs	0.142 (0.003)	0.215 (0.003)	0.130 (0.003)	0.210 (0.003)	
Active chronic conditions	0.121 (0.003)	0.182 (0.003)	0.148 (0.003)	0.267 (0.003)	
Best-to-worst difference	0.044	0.033	0.043	0.126	

Table 3. Doctors’ decisions versus algorithmic predictions. For those enrolled in the high-risk care management program (1.3% of our sample), we first show the fraction of the population that is Black, as well as the fraction of all costs and chronic conditions accounted for by these observations. We also show these quantities for four alternative program enrollment rules, which we simulate in our dataset (using the holdout set when we use our experimental predictors). We first calculate the program

enrollment rate within each percentile bin of predicted risk from the original algorithm and either (i) randomly sample patients or (ii) sample those with the highest predicted number of active chronic conditions within a bin and assign them to the program. The resultant values are then compared with values obtained by simply assigning the aforementioned 1.3% of our sample with (iii) the highest predicted cost or (iv) the highest number of active chronic conditions to the program.

Population	Fraction Black (SE)	Fraction of all costs (SE)	Fraction of all active chronic conditions (SE)
Observed program enrollment (1.3%)	0.192 (0.003)	0.029 (0.001)	0.033 (0.001)
<i>Simulated alternative enrollment rules</i>			
Random, in predicted-cost bin	0.183 (0.003)	0.044 (0.002)	0.034 (0.001)
Predicted health, in predicted-cost bin	0.269 (0.003)	0.044 (0.002)	0.064 (0.002)
Highest predicted cost	0.172 (0.003)	0.100 (0.002)	0.047 (0.002)
Worst predicted health	0.292 (0.004)	0.067 (0.002)	0.076 (0.002)

in the health sector and further afield. This is because labels are often measured with errors that reflect structural inequalities (39). Within the health sector, using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-White populations (40, 41). Outside of the health arena, credit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary (2). Policing algorithms predict measured crime, which also reflects increased scrutiny of some groups (42). Hiring algorithms predict employment decisions or supervisory ratings, which are affected by race and gender biases (43). Even retail algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to increased prices as a result (44).

This mechanism of bias is particularly pernicious because it can arise from reasonable choices: Using traditional metrics of overall prediction quality, cost seemed to be an effective proxy for health yet still produced large biases. After completing the analyses described above, we contacted the algorithm manufacturer for an initial discussion of our results. In response, the manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients. This effort confirmed our results—by one measure of predictive bias calculated in their dataset, Black patients had 48,772 more active chronic conditions than White patients, conditional on risk score—illustrating how biases can indeed arise inadvertently.

To resolve the issue, we began to experiment with solutions together. As a first step, we suggested using the existing model infrastructure—sample, predictors (excluding race, as before), training process, and so forth—but changing the label: Rather than future cost, we created an index variable that combined health prediction with cost prediction. This approach reduced the number of excess active chronic conditions in Blacks, conditional on risk score, to 7758, an 84% reduction in bias. Building on these results, we are establishing an ongoing (unpaid) collaboration to convert the results of Table 3 into a better, scaled predictor of multidimensional health measures, with the goal of rolling these improvements out in a future round of algorithm development. Of course, our experience may not be typical of all algorithm developers in this sector. But because the manufacturer of the algorithm we study is widely viewed as an industry leader in data and analytics, we are hopeful that this endeavor will prompt other manufacturers to implement similar fixes.

These results suggest that label biases are fixable. Changing the procedures by which we fit algorithms (for instance, by using a new statistical technique for decorrelating predic-

- tors with race or other similar solutions) is not required. Rather, we must change the data we feed the algorithm—specifically, the labels we give it. Producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment. But there is precedent for all of these functions in the literature and, more concretely, in the private companies that invest heavily in developing new and improved labels to predict factors such as consumer behavior (45). In addition, although health—as well as criminal justice, employment, and other socially important areas—presents substantial challenges to measurement, the importance of these sectors emphasizes the value of investing in such research. Because labels are the key determinant of both predictive quality and predictive bias, careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risks.
- REFERENCES AND NOTES**
- J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
 - S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 671 (2016).
 - A. Chouldechova, A. Roth, arXiv:1810.08810 [cs.LG] (20 October 2018).
 - A. Datta, M. C. Tschantz, A. Datta, *Proc. Privacy Enhancing Technol.* **2015**, 92–112 (2015).
 - L. Sweeney, *Queue* **11**, 1–19 (2013).
 - M. Kay, C. Matuszek, S. A. Munson, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), pp. 3819–3828.
 - B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, *IEEE Trans. Inf. Forensics Security* **7**, 1789–1801 (2012).
 - J. Buolamwini, T. Gebru, in *Proceedings of the Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.
 - A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183–186 (2017).
 - S. Corbett-Davies, S. Goel, arXiv:1808.00023 [cs.CY] (31 July 2018).
 - M. De-Arteaga et al., arXiv:1901.09451 [cs.IR] (27 January 2019).
 - M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.
 - J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *Q. J. Econ.* **133**, 237–293 (2018).
 - C. S. Hong, A. L. Siegel, T. G. Ferris, *Issue Brief (Commonwealth Fund)* **19**, 1–19 (2014).
 - N. McCall, J. Cromwell, C. Urato, "Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)" (RTI International, 2010).
 - J. Hsu et al., *Health Aff.* **36**, 876–884 (2017).
 - L. Nelson, "Lessons from Medicare's demonstration projects on disease management and care coordination" (Working Paper 2012-01, Congressional Budget Office, 2012).
 - C. Vogeli et al., *J. Gen. Intern. Med.* **22** (suppl. 3), 391–395 (2007).
 - D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, *Health Aff.* **33**, 1123–1131 (2014).
 - J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **105**, 491–495 (2015).
 - G. Hileman, S. Steele, "Accuracy of claims-based risk scoring models" (Society of Actuaries, 2016).
 - J. Kleinberg, S. Mullainathan, M. Raghavan, arXiv:1609.05807 [cs.LG] (19 September 2016).
 - A. Chouldechova, *Big Data* **5**, 153–163 (2017).
 - V. de Groot, H. Beckerman, G. J. Lankhorst, L. M. Bouter, *J. Clin. Epidemiol.* **56**, 221–229 (2003).
 - J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, *J. Clin. Epidemiol.* **64**, 749–759 (2011).
 - A. K. Parekh, M. B. Barton, *JAMA* **303**, 1303–1304 (2010).
 - D. Ettehad et al., *Lancet* **387**, 957–967 (2016).
 - K.-T. Khaw et al., *BMJ* **322**, 15 (2001).
 - K. Fiscella, P. Franks, M. R. Gold, C. M. Clancy, *JAMA* **283**, 2579–2584 (2000).
 - N. E. Adler, K. Newman, *Health Aff.* **21**, 60–76 (2002).
 - N. E. Adler, W. T. Boyce, M. A. Chesney, S. Folkman, S. L. Syme, *JAMA* **269**, 3140–3145 (1993).
 - M. Alsan, O. Garrick, G. C. Graziani, "Does diversity matter for health? Experimental evidence from Oakland" (National Bureau of Economic Research, 2018).
 - K. Armstrong, K. L. Ravenell, S. McMurphy, M. Putt, *Am. J. Public Health* **97**, 1283–1289 (2007).
 - M. Alsan, M. Wanamaker, *Q. J. Econ.* **133**, 407–455 (2018).
 - M. van Ryn, J. Burke, *Soc. Sci. Med.* **50**, 813–828 (2000).
 - K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4296–4301 (2016).
 - J. J. Escarce, F. W. Puffer, in *Racial and Ethnic Differences in the Health of Older Americans* (National Academies Press, 1997), chap. 6; www.ncbi.nlm.nih.gov/books/NBK109841/.
 - S. Passi, S. Barocas, arXiv:1901.02547 [cs.CY] (8 January 2019).
 - S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **107**, 476–480 (2017).
 - K. E. Joynt Maddox et al., *Health Serv. Res.* **54**, 327–336 (2019).
 - K. E. Joynt Maddox, M. Reidhead, A. C. Qi, D. R. Nerenz, *JAMA Intern. Med.* **179**, 769–776 (2019).
 - K. Lum, W. Isaac, *Significance* **13**, 14–19 (2016).
 - I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," available at SSRN (2016); <https://ssrn.com/abstract=2746078>.
 - S. DellaVigna, M. Gentzkow, "Uniform pricing in US retail chains" (National Bureau of Economic Research, 2017).
 - C. A. Gomez-Uribe, N. Hunt, *ACM Trans. Manag. Inf. Syst.* **6**, 13 (2016).

ACKNOWLEDGMENTS

We thank S. Lakhtakia, Z. Li, K. Lin, and R. Mahadehwari for research assistance and D. Buefort and E. Maher for data science expertise. **Funding:** This work was supported by a grant from the National Institute for Health Care Management Foundation.

Author contributions: Z.O. and S.M. designed the study, obtained funding, and conducted the analyses. All authors contributed to reviewing findings and writing the manuscript. **Competing interests:** The analysis was completely independent: None of the authors had any contact with the algorithm's manufacturer until after it was complete. No authors received compensation, in any form, from the manufacturer or have any commercial interests in the manufacturer or competing entities or products. There were no confidentiality agreements that limited reporting of the work or its results, no material transfer agreements, no oversight in the preparation of this article (besides ethical oversight from the approving IRB, which was based at a non-profit academic health system), and no formal relationship of any kind between any of the authors and the manufacturer.

Data and materials availability: Because the data used in this analysis are protected health information, they cannot be made publicly available. We provide instead a synthetic dataset (using the R package synthpop) and all code necessary to reproduce our analyses at <https://gitlab.com/labsysmed/dissecting-bias>.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/366/6464/447/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S4
References (46–51)

8 March 2019; accepted 4 October 2019
10.1126/science.aax2342

Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines

Falaah Arif Khan
New York University
New York, NY USA
fa2161@nyu.edu

Eleni Manis
S.T.O.P.
New York, NY USA
eleni@stopspying.org

Julia Stoyanovich
New York University
New York, NY USA
stoyanovich@nyu.edu

ABSTRACT

In this work we use Equal Opportunity (EO) doctrines from political philosophy to make explicit the normative judgements embedded in different conceptions of algorithmic fairness. We contrast formal EO approaches that narrowly focus on *fair contests* at discrete decision points, with substantive EO doctrines that look at people's *fair life chances* more holistically over the course of a lifetime. We use this taxonomy to provide a moral interpretation of the impossibility results as the incompatibility between different conceptions of a *fair contest* – forward-facing versus backward-facing – when people do not have *fair life chances*. We use this result to motivate substantive conceptions of algorithmic fairness and outline two plausible *fair decision procedures* based on the luck egalitarian doctrine of EO, and Rawls's principle of fair equality of opportunity.

ACM Reference Format:

Falaah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2022. Towards Substantive Conceptions of Algorithmic Fairness: Normative Guidance from Equal Opportunity Doctrines. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '22)*, October 6–9, 2022, Arlington, VA, USA. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3551624.3555303>

1 EQUALITY OF OPPORTUNITY

Equality of Opportunity (EO) is a philosophical doctrine that objects to morally arbitrary and irrelevant factors affecting people's access to desirable positions, and the social goods attached to them (such as opportunity and wealth). In an EO-respecting society, all people, irrespective of their morally arbitrary characteristics, such as socio-economic background, gender, race, or disability status, have comparable access to the opportunities that they desire. Similarly, in fair machine learning (fair-ML), we are usually interested in ensuring that the outputs of algorithmic systems, specially those used in critical social contexts, do not systematically skew along the lines of membership in protected groups based on gender, race, or disability. In so far as protected groups are constructed on the basis of morally arbitrary factors, the moral desiderata of EO doctrines from political philosophy align exactly with the fairness-related concerns in machine learning. In this work, we employ ideas from the rich EO literature from political philosophy [2, 3, 6, 14, 15, 23, 26, 30–33]

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

EAAMO '22, October 6–9, 2022, Arlington, VA, USA
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9477-2/22/10...\$15.00
<https://doi.org/10.1145/3551624.3555303>

to clarify the normative foundations of fairness and justice-related interventions, and gauge the efficacy of current algorithmic approaches that attempt to codify these criteria.

1.1 Principles of EO

There are two broad principles of EO, namely, *the principle of fair contests* and *the principle of fair life chances*.

1.1.1 Fair contests. The principle of fair contests, commonly understood as the *nondiscrimination principle*, says that competitions for desirable positions should be open to all and should be adjudicated based on competitors' relevant merits, or qualifications. In any fair contest, the most qualified person wins. Conversely, fair contests do not judge competitors on the basis of irrelevant characteristics, especially excluding morally arbitrary factors such as gender, race, and socio-economic status that are not properly understood as qualifications at all.

The principle of fair contests has been very influential in fair-ML and has guided statistical measures and algorithmic interventions that conceptualize *fairness* as *nondiscrimination*.

1.1.2 Fair life chances. The principle of fair life chances says that people's chances of success over a lifetime should not depend on morally arbitrary factors. It takes a holistic view of equal opportunity by comparing the *opportunity sets* that people have over the course of a lifetime, and is popularly understood as a principle that *levels the playing field*.

The principle of fair life chances has been almost entirely overlooked in fair-ML, and this omission explains some of the limitations in current approaches, as we will discuss shortly.

1.2 Domains of EO

According to Fishkin [15], there are, broadly, three domains of EO:

1.2.1 Fairness at a specific decision point. The first domain comprises the discrete points at which social goods, such as employment, admissions, and loan decisions are distributed. EO doctrines compel us to think about whether outcomes of decision-making at discrete decision points are influenced by morally arbitrary factors.

1.2.2 Equality of developmental opportunities. The second domain comprises educational and other foundational opportunities that shape people's ability to compete for desirable positions in the first domain. EO doctrines are also concerned with whether people had comparable developmental opportunities to build up their qualifications ahead of competitions.

1.2.3 Equality of opportunities over the course of a lifetime. Lastly, EO doctrines also compel us to look more broadly at the *opportunity*

sets to which people have access over the course of a lifetime, and whether these bundles are comparable.

1.3 Roadmap

Different interpretations of the two principles discussed in Section 1.1, and to which domain they apply, give rise to different conceptions of EO. We fix notation in Section 2. We then discuss Formal EO doctrines—which emphasize the principle of *fair contests* at discrete decision points—in Sections 3 and 4. We go on to provide a moral interpretation of the impossibility results in fair-ML as the incompatibility between a forward-facing vs. backward-facing conception of a fair contest in Section 5. In Section 6, we discuss Substantive EO doctrines—which emphasize the principle of *fair life chances* and target all three domains—as they are classically understood. Next, in Section 7, we provide modern re-interpretations of these doctrines that are more amenable to real-world decision-making. In Section 8, we arrange the moral desiderata of different EO doctrines into a *Fairness as Equal Opportunity* taxonomy. In Section 9, we demonstrate how our EO-based framework can provide normative guidance in practical contexts using a hypothetical example of college admissions, and the real-world case study of COMPAS. We compare our framework with contemporary work in Section 10, and then conclude with a discussion about the importance of grounding algorithmic approaches in strong normative foundations, as well as the limitations in guidance that EO doctrines can provide towards this end.

2 NOTATION

Before discussing different EO doctrines and their codification in fair-ML, let us fix some notation: Algorithmic decision-making involves predicting an outcome y' given a set of observations (X, y) , where the X s are covariates/features and the y s are the targets. In a given context, we assume that covariates X can be partitioned into “morally relevant” attributes A , and “morally arbitrary” attributes (such as gender, race, age, or disability status) S . “Fair” decision-making is concerned with satisfying some moral desiderata C , with respect to the set of attributes S , to which we commonly refer as the “sensitive attributes”.

We can apply EO doctrines to predictive problems if the target y indicates a social outcome; where people receive or don’t receive some desirable social good, the covariates A measure some notion of merit, and protected-groups are constructed on the basis of a morally arbitrary and irrelevant characteristic or characteristics $s \in S$. For example, predicting the risk of loan default can be posed as the problem of allocating a positive or negative lending decision, and the co-variates measure “financial qualifications” like income, repayment behavior, or net worth. Racial discrimination is illegal in lending, and so the morally arbitrary feature set in this example would include race.

For the rest of the discussion, we restrict ourselves to decision-making of this type, where the predicted outcome y' is a real valued score used to make a distributive/allocative decision. *We stress that EO doctrines are only suitable to predictive contexts that can be posed as the problem of distribution of some social good, on the basis of some relevant qualifications/merit.* We emphasize that EO doctrines

are inapplicable to algorithmic contexts that cannot be posed as distributive problems [30].

3 FORMAL EO

Formal EO doctrines specify the moral desiderata of *fair contests*: they say that no person should be excluded from a competition for a desirable position on the basis of morally arbitrary criteria. Further, in a *fair contest*, people should only be judged on the basis of their relevant merit, and so, people with comparable relevant qualifications should get the same outcome.

Formal EO, commonly known as *careers open to talents*, is only concerned with the first domain of equal opportunity—fairness at a discrete decision point [32]. Formal EO is not attentive to whether people had comparable access to developmental opportunities to build qualifications leading up to the *fair contest* (the second domain of EO), nor whether people will have comparable opportunity sets over the course of their lifetimes (the third domain of EO).

In Bernard Williams’s famous example of a warrior society, formal EO is achieved when warrior positions are open to all, and all are allowed to compete—not just the children of warrior parents [32]. However, formal EO fails to prevent privilege from being converted into qualifications in advance of the competition, whereby children from non-warrior families have no realistic chance of winning the contest without the resources and training that is afforded to children of warrior parents.

3.1 Formal EO as Fairness Through Blindness

Decision-making that is blind to irrelevant characteristics is consistent with formal EO. In fair-ML, a prominent codification of formal EO is *fairness through blindness* [13], where protected (and morally irrelevant) attributes are removed from the data, and a group-blind classifier is produced. To the extent that irrelevant characteristics (and their proxies) can be successfully excluded from an algorithm’s pipeline, formal EO can make progress toward its aim of rejecting the use of morally irrelevant features as the basis for awarding privileged outcomes. In practice, however, formal EO is often too weak to enhance fairness, as has been demonstrated in both the digital and analog age [1, 27].

Take the example of the U.S. “Ban the Box” campaign, which was aimed at passing legislation that required employers to be *blind* to candidates’ criminal histories during initial assessments of qualifications¹. The campaign was aimed at eliminating the check box on job applications that asked applicants to indicate whether they have a criminal history. Excluding criminal history from initial screenings of candidates captures formal EO’s conception of a fair contest because it attempts to ensure that justice-involved persons are judged fairly on the basis of their qualifications and not dismissed out of hand. However, this *formally fair* policy ended up having the opposite effect in practice. Field studies showed that in the absence of individual information about applicants’ criminal histories, employers end up making group-level assumptions about prior criminal justice involvement [1]. This meant that applicants with no justice involvement who belonged to groups with higher (perceived) conviction rates, such as young black males, were adversely affected,

¹<https://bantheboxcampaign.org>

while white applicants with criminal justice involvement received the benefit of the doubt.

Similar statistical discrimination due to the exclusion of group-level information is seen when formal EO is encoded into algorithmic decision-making systems. For example, Lipton et al. [27] demonstrate a “gender-blind” algorithm that discriminates on the basis of “inferred” gender at the group level when gender information at the individual level is excluded. As a result, the algorithm adversely treats applicants that it perceives as women (including men with long hair) and favors candidates that it infers to be men (including women with short hair).

The limitations of fairness through blindness are well-appreciated in fair-ML [13], so we instead turn to an alternative, stronger conception of formal EO.

3.2 Formal EO as Calibration

A well-calibrated test satisfies formal EO’s conception of a fair contest because it ensures that the likelihood of getting a positive outcome does not depend upon morally arbitrary group membership ($s \in S$):

$$P(y = 1|y' = c, s = 0) = P(y = 1|y' = c, s = 1).$$

Put differently, if two individuals have the same predicted score y' (relevant merit) and only differ on group membership s (morally irrelevant factors) then they are likely to get the same outcome from a well-calibrated test.

3.3 Formal EO as Predictive Parity

A test that satisfies predictive parity at threshold p is formal-EO compliant because it ensures that the likelihood of getting a positive outcome is the same for all high-performing individuals, irrespective of morally arbitrary group membership ($s \in S$):

$$P(y = 1|y' > p, s = 0) = P(y = 1|y' > p, s = 1).$$

Intuitively, formal EO mandates that all people who have job-relevant qualifications ($y' > p$) should have the same chance of receiving a positive outcome ($y = 1$), irrespective of their irrelevant, morally arbitrary attributes (s), and predictive value parity as a fairness criterion reflects exactly this.

4 FORMAL-PLUS EO

The strength of formal EO as a moral framework to design fair contests relies greatly on the ability to correctly measure candidates’ relevant merit. In the codifications of formal EO as calibration and predictive parity, we are making an assumption about the predicted score y' , namely, that it does, in fact, measure the applicant’s “relevant merit.” This is a strong assumption and one that does not hold in societies with historic systemic inequality. Fishkin [15] writes: “When the formal egalitarian argued that the warrior children have more merit than the non-warrior children, that view depended on a factual premise: that the warrior test did what it was designed to do and accurately predicted future warrior performance. What if it did not?”

For example, think of the SAT as a predictor of college success: When students can afford to do a lot of preparation, scores are an inflated reflection of their college potential. On the other hand, when students have limited access to preparatory material, the

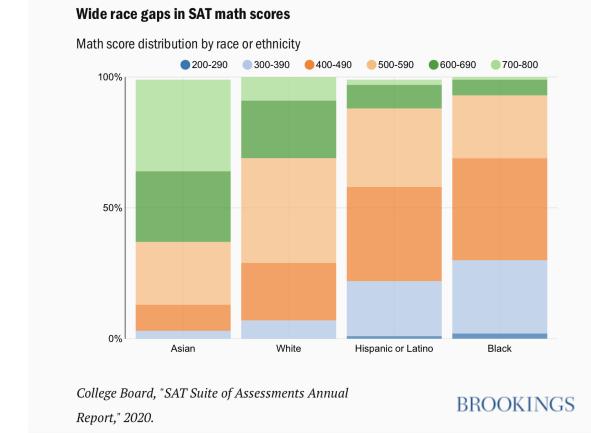


Figure 1: Distribution of SAT math scores by race or ethnicity

SAT underestimates their college potential. The SAT systematically over-predicts the future performance of more privileged students, while systematically under-predicting future performance of less privileged students: that is, the test’s validity as a predictor of college potential varies across groups. Such disparity in standardized test scores along the lines of race (and gender) has been observed in real-world contexts and is shown in Figure 1, reproduced from a Brookings 2020 report.²

In order to correct for this circular problem, where the measurement of relevant merit itself tracks morally irrelevant privilege and disprivilege, Fishkin [15] proposes a version of formal EO that he calls “formal-plus,” which adjusts test results for members of groups that are systematically underestimated by a test. From a formal-plus perspective, a *fair contest* is one in which test errors do not track (morally arbitrary) group membership.

4.1 Formal-plus EO as Error Rate Balance

A test with balanced error rates at a threshold p captures formal-plus EO’s conception of a fair contest because it ensures that test performance (i.e., false-positive rate and false-negative rate) does not skew with morally irrelevant group membership ($s \in S$):

$$P(y' > p|y = 0, s = 0) = P(y' > p|y = 1, s = 1) \text{ and}$$

$$P(y' <= p|y = 1, s = 0) = P(y' <= p|y = 0, s = 1)$$

4.2 Formal-plus EO as Equalized Odds

Fishkin [15] further explains that, in the absence of a “perfectly accurate test,” and with the understanding of which groups the test tends to underestimate, the formal-plus EO conception of a *fair contest* would “give compensatory bonus points” on the test to those whose future performance the test itself predictably underestimates. The idea of compensatory bonus points is simply to “make more accurate predictions about who, in the future, will actually be the best warriors.” The “equal opportunity” measure and algorithm from Hardt et al. [18] exactly captures formal-plus EO’s

²<https://www.brookings.edu/blog/up-front/2020/12/01/sat-math-scores-mirror-and-maintain-racial-inequity/>

moral desiderata, because it measures unfairness as the disparity in true-positive rates between groups, and randomly assigns positive outcomes when predicted scores fall between group-specific thresholds, to achieve the desired parity.

5 IMPOSSIBILITY RESULTS: FAIR CONTESTS WITHOUT FAIR LIFE CHANCES

Concurrent work by Chouldechova [9] and Kleinberg et al. [25] showed that it is impossible to simultaneously achieve parity in error rates and positive predictive value for different groups, if the prevalence (or base rates) differs between these groups. We will now provide a moral interpretation of these impossibility results, through the lens of the EO doctrines that we have discussed so far, as the incompatibility between two different conceptions of a *fair contest* – one that seeks to fairly reward *past* performance, versus one that seeks to fairly estimate *future* performance – when people do not have *fair life chances*. We use this result to motivate the need for substantive conceptions of algorithmic fairness, which we discuss in the rest of the paper.

Formal EO's conception of a fair contest takes a moral *end-point* view [23] – one that rewards the qualifications that people have already developed. Formal EO codified as predictive parity mandates that whoever possesses the relevant qualification score $y' > p$ (estimated from past performance) should be given the positive outcome $y = 1$, irrespective of morally arbitrary group membership s . Kleinberg et al. [25]'s discussion of calibration as a fairness criterion is strikingly similar to the moral desiderata of formal EO. They write: “This [calibration] means we are justified in treating people with the same score comparably with respect to the outcome, rather than treating people with the same score differently based on the group they belong to.” This is exactly an end-point view of a fair contest – one that rewards people with a comparable qualification score (computed on existing merit, judged by past performance) comparably.

On the other hand, formal-plus EO's conception of a fair contest takes a humane *starting-point* view [23] – one that aims to correctly estimate people's likelihood of succeeding in the position on offer. Once again, Kleinberg et al. [25]'s discussion of balance for the positive and negative class as a fairness criterion is strikingly similar to Fishkin [15]'s conception of formal-plus EO: “The second [balance for the negative class] and the third [balance for the positive class] ask that if two individuals in different groups exhibit comparable *future behavior* (negative or positive), they should be treated comparably by the procedure.” This is exactly formal-plus EO's starting-point view of a fair contest – one that treats people with comparable estimated future performance comparably.

Conversely, formal-plus EO would object to a test with unequal error rates across groups, because its estimate of people's future performance skews along the lines of morally irrelevant privilege/disprivilege. This is mirrored in the discussion from Kleinberg et al. [25]: “In other words, a violation of, say, the second condition [balance for the negative class] would correspond to the members of the negative class in one group receiving consistently higher scores than the members of the negative class in the other group, despite the fact that the members of the negative class in the higher-scoring group have done nothing to warrant these higher scores.”

The prevalence or base rates in a particular population is the fraction of people who possess a certain quality/qualification or receive a positive outcome (based on the existence of that quality) [9, 25]. For example, in the context of hiring, the base rate in the female population is the fraction of female candidates who receive a positive outcome—a hiring offer—among all the female applicants. Fair life chances is a broad philosophical concept, but in the context of a discrete decision, we can think of base rates among populations as a proxy for their life chances. For example, if there was gender-equality in the workforce, and women had the same employment prospects as men, then we would expect equal proportions of women and men to receive a positive hiring outcome.

Putting this together, a philosophical interpretation of the impossibility results, through the lens of EO doctrines, says that it is impossible to design a *fair contest* that simultaneously rewards people's past qualifications and accurately estimates people's future prospects of success, if people did not have fair (comparable) life chances. This is simply because morally arbitrary and irrelevant factors weigh heavily on people's achievement—as evidenced both in past performance and in the estimation of future performance.

The empirical results in critical domains such as criminal justice that led to the impossibility results in fair-ML [10, 25] are a stark demonstration of this fact: that people do not have comparable life chances and that morally arbitrary characteristics such as gender and race do weigh heavily on people's prospects of success. Importantly, through the lens of EO doctrines, we can see the limitations of our current approaches in their narrow focus on designing fair contests at discrete decision points. We now discuss substantive EO doctrines, to pivot future directions of fair-ML research towards substantive conceptions of algorithmic fairness.

6 SUBSTANTIVE EO

In order to design interventions that improve people's life chances we need substantive EO. Substantive EO focuses on giving people the opportunities to substantively build up their qualifications, so that when they do go on to compete for desirable social positions, they truly have a chance of winning. These qualifications-building opportunities fall into the second domain of EO, discussed in Section 1.2, and constitute *developmental opportunities*. The motivating reason behind both equality of developmental opportunities and EO over a lifetime is that morally arbitrary circumstances of birth, such as warrior parentage, should not determine people's *life* prospects.

There are several conceptions of substantive EO, but in this paper we will limit our discussion to two highly influential doctrines that are relevant to fair-ML, namely, Rawls's fair EO [30] and luck egalitarian EO [14, 31].

6.1 Limitations of Formal Doctrines

Before we move on to substantive EO, let us look at why formal EO doctrines fall short of satisfying all of our fairness-related concerns. Arbitrary and morally irrelevant privileges weigh heavily on the outcomes of formally *fair* competitions because people can leverage them to build qualifications in advance of competitions. This undermines the promise of formal doctrines—that only relevant skill will be rewarded—because it fails to stop gains from being distributed along the lines of privilege and disprivilege. We call

this the *before* problem: before competitions, people are allowed to exercise their privilege to develop relevant qualifications.

Formal doctrines also have an “after” problem. After formally *fair* competitions, winners are set up for even more success. A candidate that is hired for a job is consequently granted access to more training and job experience. This makes them even more competitive in the next competition for jobs. Conversely, those who lose at first, lose opportunities for skill development, leading to more losses. Formally fair competitions create a snowball effect, where early-on winners get further ahead while early-on losers fall further behind.

Formal EO’s “before” and “after” problems compound, with privileged candidates using their competitive advantages to win early on, thus securing more developmental opportunities, which enable further wins. Anderson [2] calls this phenomenon “discrimination laundering.” Formal doctrines cannot adequately address these problems, because they are limited to measuring people’s qualifications accurately, and to excluding irrelevant information.

To meaningfully correct for social inequalities we need substantive EO doctrines. Fishkin [15] writes: “The reason that the warrior society is interesting is that, per stipulation, it is not simply the case that children of warriors appear, through test-related artifice, most likely to be the best future warriors. The point of the example is that the children of warriors really are the most likely to grow into the best adult warriors as a result of their accumulated childhood advantages.” This is exactly the target of substantive EO doctrines: making sure that people have comparable “opportunity sets” over the course of a lifetime.

6.2 Rawls’s Fair EO

Rawls’s principle of fair equality of opportunity (FEO) says that *equally talented people should have equal prospects of success*. Rawls [30] writes: “Assuming that there is a distribution of natural assets, those who are at the same level of talent and ability, and have the same willingness to use them, should have the same prospects of success regardless of their initial place in the social system.”

In setting out his theory of justice that embeds the FEO principle, Rawls identifies two distributive mechanisms: the “social lottery,” which distributes people their initial positions in the social system, and the “natural lottery,” which distributes people their native talent and ability. He writes [30]: “We do not deserve our place in the distribution of native endowments, any more than we deserve our initial starting place in society.” Rawls posits that the natural and social lotteries are not by themselves unjust, but it is the way that institutions have been set up that leads to inequality along the lines of morally arbitrary characteristics. His principles of justice are designed to help a society appropriately mitigate their effects.

With this in mind, Rawls’s theory of justice [30] posits the following principles that would regulate the distribution of primary social goods (including wealth and opportunity) by institutions in a just society:

- (1) Rights and liberties: Everyone has the same inalienable right to equal basic liberties.
- (2)(a) Principle of Fair EO: All offices and positions must be open to all under conditions of fair equality of opportunity.

- (b) Difference principle: Any social inequality must be applied in such a manner that they be of the greatest benefit to the least advantaged.

The principles are lexically ordered, in that people’s basic rights and liberties cannot be infringed upon while bringing about FEO, nor can the Difference Principle be applied in a way that violates FEO. People’s fundamental rights are given highest priority. Next, the principle of FEO regulates the distribution of desirable social positions such that people benefit from their arbitrary endowments from the natural lottery, but are not disadvantaged by the social lottery. Once this has been satisfied, the Difference Principle is applied, seeking to redistribute social inequality to the greatest benefit of the worst-off group, so as to limit the current and *future* effect of both lotteries.

Rawls’s principles, including his principle of fair EO, are regulatory and holistic in nature—they are to be applied iteratively to regulate the distribution of social goods by institutions. It is unclear how to port these principles, as currently understood, to discrete decision-making contexts—the kind that are the focus of fair-ML. In Section 7 we will propose a modern re-interpretation that allows for such a translation.

6.3 Luck Egalitarian EO

Luck egalitarian EO levels the playing field by making competitors’ opportunities comparable, and then allows individual choices and effort to determine the outcomes of competitions. Any resulting disparity in outcomes is morally acceptable because it is due to differential individual effort, not differential fortune.

At a discrete decision point (the first domain of EO), morally arbitrary circumstances have already weighed heavily on people’s abilities. The luck egalitarian conception of a fair contest partitions a person’s qualifications into two sets—matters of “option luck” or “choice luck” for which it is morally correct to hold the individual accountable, and effects of “brute luck” that are morally irrelevant. Luck egalitarian EO says that people’s outcomes (access to desirable positions) should only be affected by the former, and no matters of brute luck should affect the outcome of a fair contest.

The hard question now is how to make this correction. How do we separate the effects of brute luck (circumstance) from the effects of responsible choices (effort)? Roemer [31] proposed a version of EO that fulfills the moral desiderata of the luck egalitarian doctrine while bypassing the need to make an explicit separation between “responsible effort” and “arbitrary circumstance”. Instead, Roemer introduced the idea of “types”: people with the same morally arbitrary circumstance are of the same “type.” Now, for a certain matter of arbitrary circumstance (e.g., family income), the entire population can be partitioned into types (e.g., “high income,” “medium income,” and “low income”). Using this idea of circumstance-types, Roemer posits that, in comparing the effort of candidates, we should correct for the fact that those efforts are *drawn from different distributions*. In other words, effort distributions are characteristic of the type, and not of the individual, and this difference is due to a morally arbitrary factor of circumstance, for which individuals should not be held accountable. Now, in evaluating an individual’s qualifications (effort), we should only compare them to others of the same type (with the same circumstance).

For example, in Figure 1, we see a large disparity in the SAT score distributions broken down by race or ethnicity. In a luck egalitarian procedure, we would evaluate students' test scores based on where they placed within the score distribution of their particular race/ethnicity. From a moral standpoint, two individuals of different types are equally qualified for a desirable position if they lie at the same quantile of the effort distribution of their type. Hence, Roemer's conception of the a *fair contest* evaluates people by correcting for the unequal (unfair) life chances they've had in the past by ranking them in their effort-type distribution.

7 A MODERN RE-INTERPRETATION OF SUBSTANTIVE EO DOCTRINES

Substantive EO doctrines, discussed in Section 6, have much stronger moral desiderata than formal EO doctrines, discussed in Sections 3 and 4, but they also have severe shortcomings, which can preclude applying them to real-world contexts that are the focus of fair-ML. Rawls's theory has been the subject of much debate and criticism for being limited to ideal theorizing: There is very little guidance from Rawls about how to apply his principles of justice in practice, or how to bring about FEO in a world where people do not have comparable life prospects. Luck egalitarianism, specially strict interpretations of the doctrine, also has severe shortcomings such as issues of agency and autonomy in holding people responsible for certain types of luck and not for others.

In this section we provide a modern interpretation of luck egalitarian EO and Rawls's FEO, in a way that is both consistent with the original doctrines and amenable to providing normative guidance in practical contexts: We classify the luck egalitarian approach as a *backward-facing*, indirect approach to equalizing people's opportunity sets, and Rawls's as a *forward-facing*, direct approach to bringing about substantive EO.

Table 1 summarizes our classification of the normative approaches of different EO doctrines, and we elaborate on this in the remainder of the section.

Table 1: Classification of EO doctrines

	Backward-facing	Forward-facing
Fair contests	Formal	Formal-plus
Fair life chances	Luck egalitarian	Rawls

7.1 Luck Egalitarian EO as a Backward-Facing View of Fair Life Chances

The luck egalitarian view acknowledges that differences in people's qualifications at the point of competition are, at least in part, due to morally arbitrary circumstances (matters of brute luck) and so a fair competition should only evaluate candidates on the basis of their propensity to expend effort, and not on the qualifications that are built from this effort. Intuitively, the idea is that people who are disadvantaged by circumstances will have to put in far greater effort to reach the same level of ability as compared to people with advantageous circumstances, and so effort is the correct rubric of achievement, not ability.

From a practical standpoint, the luck egalitarian approach gives rise to a two-step procedure of substantive EO: first, control for people's unequal life chances, and then conduct fair contests on the basis of these adjusted qualifications. We interpret luck egalitarianism as a *backward-facing* conception: it corrects for *past* effects of brute luck, and then allows individual effort to decide future outcomes. It does improve people's life chances by distributing opportunities to which they likely would not have had access, had we not corrected for their unequal life chances in the past. However, this conception does not correct for the differential effort that will be required by people with different circumstances to excel in that position in the *future*.

7.2 Rawls's FEO as a Forward-Facing View of Fair Life Chances

Rawls's principle of fair equality of opportunity (FEO) says that *equally talented people should have equal prospects of success*. Let us unpack the principle into two components: the first part deals with identifying "equally talented" people, whereas the second part says that outcomes should be distributed in such a way that gives these "equally talented" people "equal prospects of success."

An implementation of the first part of the principle aligns with the luck egalitarian approach: we adjust our measurement of people's abilities after controlling for the effects of the social lottery in order to approximate their "native talent." The second part of the principle is where the two conceptions diverge: the luck egalitarian stops after correcting for *past* effects and simply distributes outcomes based on this corrected measurement. Rawls goes one step further, and distributes outcomes in a way that also makes people's *future* prospects of success (i.e., their prospects of succeeding in the next contest) comparable. Hence, we interpret Rawls's FEO as a *forward-facing* view of the principle of fair life chances.

We summarize our interpretation of EO doctrines in the next section, and go on to illustrate the distinction between practical applications of different EO doctrines using examples in Section 9.

8 FAIRNESS AS EQUAL OPPORTUNITY TAXONOMY

We summarize the moral desiderata and normative approaches of different EO doctrines in the *Fairness as Equal Opportunity* taxonomy given in Table 2. This gives us guidance about what value judgements our fairness interventions codify, and helps us design a suitable fairness intervention for a given context based on our normative judgements. For example, we can decide that outcomes correspond to rewards for past performance and choose a *formal* approach. Or, we can decide that the decision-making context requires selecting people who are most likely to succeed in the future and take a *formal-plus* approach instead. We can apply a two-step substantive approach following the luck egalitarian view by first adjusting the measurement of people's qualifications to correct for past effects of morally arbitrary factors, and then apply any suitable selection procedure on these adjusted qualification scores. Lastly, we can gauge that the decision-making context is a critical developmental opportunity, and choose Rawls's approach that forgoes maximum utility today in favor of improved equity tomorrow.

Table 2: Fairness as Equal Opportunity taxonomy

Doctrine	Moral desiderata	Normative approach
Formal	Fair contests should only measure morally relevant qualifications	Accurately measure past performance
Formal-plus	The performance of fair contests should not skew along the lines of morally irrelevant features	Accurately estimate future performance
Substantive: Luck egalitarian	Matters of brute luck should not affect people's outcomes	Distribute outcomes on the basis of effort, after correcting for the past effects of morally arbitrary circumstances
Substantive: Rawls	Equally talented people should have equal prospects of success	Distribute outcomes to equalize future prospects of success of people who have the same native talent, irrespective of arbitrary circumstance

9 NORMATIVE GUIDANCE

9.1 Illustrative Example: College Admissions

We now present an example to illustrate how different EO doctrines conceptualize a *fair contest*. Suppose, for the sake of argument, that we are making a college admissions decision based on a single standardized score, shown in Figure 2. Each applicant belongs to one of two demographic groups, A and B, where group membership is based on some morally arbitrary characteristic, say race or gender. We can see that members of group A (the distribution to the left) have systematically lower scores than members of group B (the curve to the right), and we posit that this is due to the effect of morally arbitrary circumstances that differ between the two groups, and are not due to any innate difference in talents in the two groups. The question now is: How do we distribute the desirable outcome of a positive admissions decision to people from both groups, in a way that is *fair*?

Following formal EO, we would simply decide a threshold on the score, which corresponds to the level of past performance that we find deserving of a reward. This threshold is shown in the figure as “Formal threshold”. Positive outcomes are distributed to everyone who has a score (y') that is higher than this threshold. As we can see from the figure, although we did not explicitly set this threshold based on any morally irrelevant factors (we simply decided a cut-off on the standardized test score), nonetheless, this threshold is prohibitively high and our selection procedure effectively eliminates all of group A. This is a backward-facing view of a *fair contest*: one that distributes outcomes based on people’s past performance on the standardized test.

A formal-plus conception of this procedure would posit that the standardized test overestimates the abilities of group B and underestimates the abilities of group A, and so the test is not a suitable measurement of applicants’ *future* academic performance. Alternatively, it would select different thresholds for each group, shown in the figure, to correct for the test’s error. The threshold for group A does admit some people from this group, and this is exactly Fishkin’s idea of “compensatory bonus points” for the group whose abilities the test underestimates. This is a forward-facing view of *fair contests* that distributes outcomes based on accurately estimating people’s future performance.

Next, a luck egalitarian approach following Roemer would look at people’s positions within the score distributions of their type (the 80th percentile for group A and group B are shown in Figure 2). The luck egalitarian would posit that people who are at the same percentile in the score distribution of their type have expended the same degree of effort in the past, and hence should receive similar outcomes from the procedure. This is a backward-facing view of substantive EO: one that designs fair contests by correcting for unequal life chances in the past, and adjusts people’s qualification score to only reflect their morally relevant effort.

In order to differentiate Rawls’s approach to this problem from the luck egalitarian one, let us further look at two individuals, Alice (from type A) and Bob (from type B), who sit at the same percentile of the score distribution of their type, and hence are equally “talented” according to Rawlsian view. We saw that the luck egalitarian would give them both the same outcome (i.e., a positive admission decision) because these individuals have demonstrated the same degree of effort in the past. By contrast, the Rawlsian would make a further consideration: How likely are Alice and Bob to succeed in this desirable position, respectively? Even if they both receive the positive outcome, Alice will probably have to work much harder than Bob to actually do well in the program. The effects of circumstance are such that the absolute amount of effort

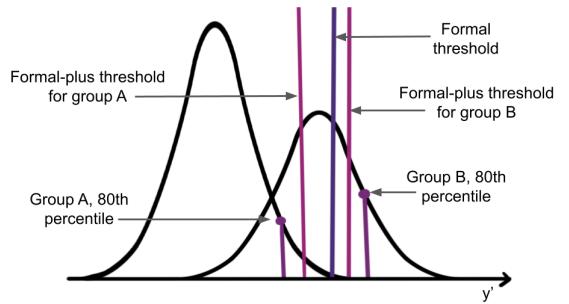


Figure 2: Distribution of test scores for groups A and B in the college admission example, discussed in Section 9.1

they have to spend to match the same level of achievement is higher for Alice than for Bob.

The luck egalitarian view does not correct for the continued future effects of circumstance, while the Rawlsian approach tries to, and makes positive admissions decisions for both Alice and Bob, but also distributes additional resources (such as tutoring and scholarships) for Alice to make up for the lack of resources and developmental opportunities they had access to, leading up to this competition. This is why we classify Rawls's FEO as a forward-facing view of a *fair contest*: it wants to set up equally talented people to have equal prospects of winning the next contest.

Note that the actual outcomes distributed by these different procedures (motivated by different EO doctrines) may, in fact, coincide in certain contexts, based on the relative effects of morally irrelevant features. What distinguishes these procedures are their normative approaches and value judgements, and not the outcomes that are finally distributed.

9.2 Illustrative Example: COMPAS

We now demonstrate how the taxonomy of EO doctrines in Table 2 enables us to debate in normative terms, based on our value judgements. Consider the infamous example of COMPAS, as exposed in an investigation by ProPublica [4]. We would like to clarify that we chose COMPAS as the example with which to illustrate our framework not because we believe it to be representative of the kinds of contexts for which algorithms *should* be designed, and *could* benefit from ethical and moral grounding. On the contrary, we unequivocally believe that an algorithm such as COMPAS *should not* be used. Yet, without grounding different statistical measures of fairness in the moral desiderata they encode we have no way to reconcile disagreements about the implications of ProPublica's findings. We view our EO-framework as a necessary first step towards being able to debate in values, and to audit algorithmic systems more holistically.

COMPAS is an algorithm that predicts the risk of violent recidivism among people awaiting trial. This can be posed as the problem of distributing access to resources such as counseling or other positive interventions, based on individuals' propensity to re-offend, and translates to a problem to which we can apply EO doctrines. Northpointe argued that COMPAS did not exhibit racial discrimination because the risk scores that it produced were equally well calibrated for both black and white defendants [12]. Connecting this argument to the moral desiderata of formal EO, we see that Northpointe's value judgement was that the algorithm "accurately" rewarded (and punished) people for their *past* actions (i.e., past criminal behavior). ProPublica, on the other hand, demonstrated that the error rates of COMPAS skewed along racial lines — the algorithm systematically overestimated the risk of black individuals and systematically underpredicted the risk of white individuals — and argued that this was evidence of racial discrimination [4]. ProPublica's critique of COMPAS can be seen through the lens of formal-plus EO: the use of COMPAS to determine parole sentences is not formal-plus EO compliant because test performance skews along the lines of race — a morally arbitrary and irrelevant feature. ProPublica appears to be making the value judgement that COMPAS ought to accurately estimate the *future* prospects of crime of

both black and white individuals. Given that the purpose of COMPAS was, in fact, to *predict recidivism* (i.e., criminal re-offense in the *future*), viewing ProPublica's audit through the lens of formal-plus EO makes their results even more compelling.

10 RELATED WORK

Equality of opportunity doctrines have been quite influential in fair-ML. Heidari et al. [20] were the first to formalize these ideas in fair-ML using economic models of EO. We were inspired by their work, but, in constructing this taxonomy using EO doctrines from political philosophy, find critical mistakes in their framework. Most importantly, their framework takes a reductive view of substantive EO doctrines, as independent fairness criteria at discrete decision points. This misrepresents the nature of substantive EO, which is not limited to the first domain of EO and takes a more holistic view of *fair contests* in regards to people's *fair life chances*, as we explain in Sections 1 and 6.

With this clarification, we refute Heidari et al. [20]'s mappings of substantive EO doctrines with statistical measures, including their assertion that statistical parity, equalized odds, and accuracy map to Rawls's FEO. We agree with their characterization of Roemer's EO, and their application of it to predictive contexts. An important point here is that formal EO and Rawls's FEO are doctrines from political philosophy, and not economics, while Roemer's EO is an economic doctrine. Hence, Heidari et al. [20]'s choice to use economic models of EO might be the cause of disagreement between our framework (using EO doctrines from political philosophy) and theirs (using economic models of EO). Specifically, Heidari et al. [20] characterize methods that are consistent with luck egalitarian EO as taking a relative view of effort, and they characterize methods that are consistent with Rawls's FEO as taking an absolute view of effort. The latter is a misconception, in that even economic interpretations of Rawls's FEO do not characterize it as a doctrine that takes an absolute view of effort [26].

Further, following Arneson [6], libertarianism has been introduced as a possible version of EO in fair-ML [20]. However, the libertarian view focuses on a narrow notion of procedural fairness: It would object to a procedure that allows illegal or unfair means of gaining access to opportunities. While libertarianism (as a limited notion of procedural fairness) may be interpreted as a fairness-preserving position from a legal standpoint, it does not satisfy EO's characteristic commitment to eliminating irrelevant and arbitrary barriers to achievement. The libertarian principle of self-ownership asserts that people are entitled to the full benefit of their natural personal endowments [26], and so the disparity in people's access to desirable positions arises simply from them exercising free will. According to this view, nothing is morally arbitrary or irrelevant, and so nothing needs to be corrected for. Hence, we reject the characterization of libertarianism as a form of EO by Arneson [6], and its adoption in fair-ML by Heidari et al. [20].

There are several contemporary works that attempt to clarify the normative foundations of algorithmic fairness [5, 7, 11, 16, 17, 21]. To the best of our knowledge, we are the first to arrange fairness desiderata under the two EO principles of *fair contests* and *fair life chances*, to introduce Fishkin [15]'s doctrine of formal-plus EO as a criterion on balance of error rates between groups, and,

most importantly, to introduce a temporal dimension to the moral desiderata of EO doctrines (forward-facing vs. backward-facing).

There has also been some interest to design fairness-enhancing interventions that go beyond a single-point view of fairness [8, 19, 22, 24, 28, 29], and we hope that our framework will help ground such work in strong normative foundations.

11 DISCUSSION

It is widely accepted that *fairness* is not a statistical concept, but rather a philosophical and moral one. Yet, current approaches in fair-ML do not explicitly state the value judgements they make. In this work, we attempt to fix this deficit by making connections between influential results in algorithmic fairness, and the normative considerations of EO doctrines. We construct a taxonomy, summarized in Table 2, that introduces a temporal dimension to influential EO doctrines (forward vs backward-facing).

In making these connections between EO doctrines and algorithmic fairness approaches, we identified limitations in current approaches, specifically, a narrow focus on designing *fair contests* at discrete decision points, without broader considerations of *fair life chances* and the overall opportunity-sets available to people. In order to move beyond formal approaches, we propose modern interpretations and plausible procedures for two substantive EO doctrines in Section 7.

It is also widely accepted that *fairness* is inherently context-specific, yet, to the best of our knowledge, an understanding of the suitability of different fairness conceptions in different contexts still lacks. We take an important step in this direction: Equal opportunity doctrines, in contrast to equality of outcome doctrines, allow us to connect the nature of the opportunity with fairness desiderata. Our taxonomy makes explicit the value judgements that go into different conceptions of a *fair contest*, underscoring that different conceptions are suitable for different contexts: Do we seek to reward past performance (formal) or accurately estimate future performance (formal-plus)? Do we care about removing the effects of past inequality (luck egalitarian), or to preemptively correct for inequality that, if left unchecked, will compound in the future (Rawls's)? Thinking in these terms can guide decision-makers in selecting a suitable fairness-related intervention that aligns with their value judgements for what is suitable for their specific context.

Lastly, we would like to re-iterate that, While the EO principles are a helpful frame within which to reason about our justice-related goals, EO doctrines offer an incomplete normative palette for thinking about discrimination. An important limitation of EO doctrines is that they are only applicable to contexts where desirable outcomes are distributed on the basis of some relevant qualification.

12 CONCLUSION

In this work we showed that extant approaches to algorithmic fairness have mainly been limited to formal conceptions of *fair contests* at discrete decision points. Through the lens of EO doctrines, we provided a moral interpretation of the impossibility results as the incompatibility between two different conceptions of a *fair contest*—a forward-facing view vs. backward-facing one—when people do not have *fair life chances*. We used this result to motivate the need for substantive conceptions of algorithmic fairness, which look

more holistically at the opportunity sets that people have available to them over the course of a lifetime, and outlined two plausible procedures to do this. We hope that our work will foster similar approaches from law and social sciences in grounding current and future research in fair-ML in strong normative foundations.

ACKNOWLEDGMENTS

This research was supported in part by National Science Foundation awards No. 1934464, 1922658, and 1916505.

REFERENCES

- [1] Amanda Agan and Sonja Starr. 2018. Ban the box, criminal records, and racial discrimination: A field experiment. *The Quarterly Journal of Economics* 133, 1 (2018), 191–235.
- [2] Elizabeth Anderson. 2010. *The Imperative of Integration*. Princeton University Press.
- [3] Elizabeth S. Anderson. 1999. What Is the Point of Equality? *Ethics* 109, 2 (1999), 287–337. <https://doi.org/10.1086/233897>
- [4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. *ProPublica* (2016).
- [5] Falah Arif Khan, Eleni Manis, and Julia Stoyanovich. 2021. Translation tutorial: Fairness and Friends. In *2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery.
- [6] Richard J. Arneson. 2018. Four Conceptions of Equal Opportunity. *Wiley-Blackwell Economic Journal* (2018). <https://doi.org/10.1111/ecoj.12531>
- [7] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA (Proceedings of Machine Learning Research, Vol. 81)*, Sorelle A. Friedler and Christo Wilson (Eds.). PMLR, 149–159. <http://proceedings.mlr.press/v81/binns18a.html>
- [8] Avrim Blum, Kevin Stangl, and Ali Vakilian. 2022. Multi Stage Screening: Enforcing Fairness and Maximizing Efficiency in a Pre-Existing Pipeline. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) (FACCT '22). Association for Computing Machinery, New York, NY, USA, 1178–1193. <https://doi.org/10.1145/3531146.3533178>
- [9] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (2017), 153–163. <https://doi.org/10.1089/big.2016.0047>
- [10] Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM* 63, 5 (2020), 82–89. <https://doi.org/10.1145/3376898>
- [11] Kathleen Creel and Deborah Hellman. 2021. The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision Making Systems. In *FACCT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, Madeleine Clare Elish, William Isaac, and Richard S. Zemel (Eds.). ACM, 816. <https://doi.org/10.1145/3442188.3445942>
- [12] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity. *Northpointe Inc. Research Department* (2016).
- [13] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. 2012. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, Shafi Goldwasser (Ed.). ACM, 214–226. <https://doi.org/10.1145/2090236.2090255>
- [14] Ronald Dworkin. 1981. What is Equality? Part 1: Equality of Welfare. *Philosophy and Public Affairs* 10, 3 (1981), 185–246. <http://www.jstor.org/stable/2264894>
- [15] Joseph Fishkin. 2014. *Bottlenecks: A New Theory of Equal Opportunity*. Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780199812141.001.0001>
- [16] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. On the (im)possibility of fairness. *CoRR* abs/1609.07236 (2016). arXiv:1609.07236 <http://arxiv.org/abs/1609.07236>
- [17] Ben Green. 2021. Escaping the "Impossibility of Fairness": From Formal to Substantive Algorithmic Fairness. *arXiv preprint arXiv:2107.04642* (2021).
- [18] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.), 3315–3323. <https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html>
- [19] Hoda Heidari and Jon Kleinberg. 2021. Allocating Opportunities in a Dynamic Model of Intergenerational Mobility. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FACCT

- '21). Association for Computing Machinery, New York, NY, USA, 15–25. <https://doi.org/10.1145/3442188.3445867>
- [20] Hoda Heidari, Michele Loi, Krishna P. Gummadi, and Andreas Krause. 2019. A Moral Framework for Understanding Fair ML through Economic Models of Equality of Opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*. ACM, 181–190. <https://doi.org/10.1145/3287560.3287584>
- [21] Corinna Hertweck, Christoph Heitz, and Michele Loi. 2021. On the Moral Justification of Statistical Parity. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) (FAccT '21). Association for Computing Machinery, New York, NY, USA, 747–757. <https://doi.org/10.1145/3442188.3445936>
- [22] Corinna Hertweck and Tim Ráz. 2022. Gradual (In)Compatibility of Fairness Criteria. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelfth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 11926–11934. <https://ojs.aaai.org/index.php/AAAI/article/view/21450>
- [23] Christopher Jencks. 1988. Whom Must We Treat Equally for Educational Opportunity to be Equal? *Ethics* 98, 3 (1988), 518–533. <http://www.jstor.org/stable/2380965>
- [24] Sampath Kannan, Aaron Roth, and Juba Ziani. 2019. Downstream Effects of Affirmative Action. In *2019 ACM Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) (FAT* '19). Association for Computing Machinery, New York, NY, USA, 240–248. <https://doi.org/10.1145/3287560.3287578>
- [25] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA (LIPIcs, Vol. 67)*, Christos H. Papadimitriou (Ed.). Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 43:1–43:23. <https://doi.org/10.4230/LIPIcs.ITCS.2017.43>
- [26] Arnaud Lefranc, Nicolas Pistolesi, and Alain Trannoy. 2009. Equality of opportunity and luck: Definitions and testable conditions, with an application to income in France. *Journal of Public Economics* 93, 11 (2009), 1189–1207. <https://doi.org/10.1016/j.jpubeco.2009.07.008>
- [27] Zachary C. Lipton, Julian J. McAuley, and Alexandra Chouldechova. 2018. Does mitigating ML's impact disparity require treatment disparity?. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.), 8136–8146. <https://proceedings.neurips.cc/paper/2018/hash/8e0384779e58ce2af40eb365b318cc32-Abstract.html>
- [28] David Liu, Zohair Shafi, William Fleisher, Tina Eliassi-Rad, and Scott Alfeld. 2021. RAWLSNET: Altering Bayesian Networks to Encode Rawlsian Fair Equality of Opportunity. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, Marion Fourcade, Benjamin Kuipers, Seth Lazar, and Deirdre K. Mulligan (Eds.). ACM, 745–755. <https://doi.org/10.1145/3461702.3462618>
- [29] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. 2018. Delayed Impact of Fair Machine Learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 3156–3164. <http://proceedings.mlr.press/v80/liu18c.html>
- [30] John Rawls. 1971. *A Theory of Justice*. Harvard University Press. <http://www.jstor.org/stable/j.ctv1f9z6v>
- [31] John E. Roemer. 2002. Equality of opportunity: A progress report. *Social Choice and Welfare* 19, 2 (2002), 455–471. <http://www.jstor.org/stable/41106460>
- [32] Bernard Williams. 1973. *The idea of equality*. Cambridge University Press, 230–249. <https://doi.org/10.1017/CBO9780511621253.016>
- [33] H. Peyton Young. 1994. *Equity: In Theory and Practice*. Princeton University Press. <https://books.google.co.in/books?id=XVK5AAAAIAAJ>

Additional Reading

This section contains references to additional reading. This reading is not required, and I do not expect you to be familiar with this material for homework assignments, quizzes, or the exam.

Fairness and machine learning: Limitations and opportunities by Solon Barocas, Moritz Hardt and Arvind Narayanan

How to use this text: This is an online textbook, you may use it for reference on any of the topics we covered during weeks 1–4. It has accompanying video tutorials by the authors, and is an excellent supplementary resource.

Fairness through Awareness by Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, Richard S. Zemel, *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, ITCS 2012*

How to use this text: This is one of the first papers on algorithmic fairness, and certainly the best-known early work on fairness in classification. We discussed this paper during week 2. You may find it helpful to skim this paper to better understand the high-level presentation in the slides, or to get into the technical details.

Abstract: We study fairness in classification, where individuals are classified, e.g., admitted to a university, and the goal is to prevent discrimination against individuals based on their membership in some group, while maintaining utility for the classifier (the university). The main conceptual contribution of this paper is a framework for fair classification comprising (1) a (hypothetical) task-specific metric for determining the degree to which individuals are similar with respect to the classification task at hand; (2) an algorithm for maximizing utility subject to the fairness

constraint, that similar individuals are treated similarly. We also present an adaptation of our approach to achieve the complementary goal of "fair affirmative action," which guarantees statistical parity (i.e., the demographics of the set of individuals receiving any classification are the same as the demographics of the underlying population), while treating similar individuals as similarly as possible. Finally, we discuss the relationship of fairness to privacy: when fairness implies privacy, and how tools developed in the context of differential privacy may be applied to fairness.

Learning Fair Representations by Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, Cynthia Dwork, *Proceedings of the 30th International Conference on Machine Learning, PMLR 2013*

How to use this text: An influential early paper on fairness in classification that follows up on Fairness through Awareness. We discussed this paper during week 2. You may find it helpful to skim this paper to better understand the high-level presentation in the slides, or to get into the technical details.

Abstract: We propose a learning algorithm for fair classification that achieves both group fairness (the proportion of members in a protected group receiving positive classification is identical to the proportion in the population as a whole), and individual fairness (similar individuals should be treated similarly). We formulate fairness as an optimization problem of finding a good representation of the data with two competing goals: to encode the data as well as possible, while simultaneously obfuscating any information about membership in the protected group. We show positive results of our algorithm relative to other known techniques, on three datasets. Moreover, we demonstrate several advantages to our approach. First, our intermediate representation can be used for other classification tasks (i.e., transfer learning is possible); secondly, we take a step toward learning a distance metric which can find important dimensions of the data for classification.

Fairness in Ranking, Part I: Score-based Ranking by Meike Zehlike, Ke Yang and Julia Stoyanovich, *ACM Computing Surveys, 2022*

How to use this text: Reach sections 1, 2 and 3 that introduce fairness in ranking, give a running example, and propose a classification framework for fair ranking methods that ties together normative and technical dimensions. Among other things, this framework illustrates how equality of opportunity (EO) doctrines can be operationalized in algorithmic fairness methods.

Abstract: In the past few years, there has been much work on incorporating fairness requirements into algorithmic rankers, with contributions coming from the data management, algorithms, information retrieval, and recommender systems communities. In this survey, we give a systematic overview of this work, offering a broad perspective that connects formalizations and algorithmic approaches across sub-fields. An important contribution of our work is in developing a common narrative around the value frameworks that motivate specific fairness-enhancing interventions in ranking. This allows us to unify the presentation of mitigation objectives and of algorithmic techniques to help meet those objectives or identify trade-offs.

In this first part of this survey, we describe four classification frameworks for fairness-enhancing interventions, along which we relate the technical methods surveyed in this article, discuss evaluation datasets, and present technical work on fairness in score-based ranking. In the second part of this survey, we present methods that incorporate fairness in supervised learning, and also give representative examples of recent work on fairness in recommendation and matchmaking systems. We also discuss evaluation frameworks for fair score-based ranking and fair learning-to-rank, and draw a set of recommendations for the evaluation of fair ranking methods.