Sociological Methods & Research
1-42

DOI: 10.1177/0049124118782533
journals.sagepub.com/home/smr
SAGE

# Fairness in Criminal Justice Risk Assessments: The State of the Art

**Richard Berk[1,2], Hoda Heidari[3], Shahin Jabbari[3], Michael Kearns[3], and Aaron Roth[3]**

## Abstract

**Objectives:** Discussions of fairness in criminal justice risk assessments typically lack conceptual precision. Rhetoric too often substitutes for careful analysis. In this article, we seek to clarify the trade-offs between different kinds of fairness and between fairness and accuracy. **Methods:** We draw on the existing literatures in criminology, computer science, and statistics to provide an integrated examination of fairness and accuracy in criminal justice risk assessments. We also provide an empirical illustration using data from arraignments. **Results:** We show that there are at least six kinds of fairness, some of which are incompatible with one another and with accuracy. **Conclusions:** Except in trivial cases, it is impossible to maximize accuracy and fairness at the same time and impossible simultaneously to satisfy all kinds of fairness. In practice, a major complication is different base rates across different legally protected groups. There is a need to consider challenging trade-offs. These lessons apply to applications well beyond criminology where assessments of risk can be used by decision makers. Examples include mortgage lending, employment, college admissions, child welfare, and medical diagnoses.

[1] Department of Statistics, University of Pennsylvania, Philadelphia, PA, USA
[2] Department of Criminology, University of Pennsylvania, Philadelphia, PA, USA
[3] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

**Corresponding Author:**
Richard Berk, Department of Statistics, University of Pennsylvania, 483 McNeil, 3718 Locust Walk, Philadelphia, PA 19104, USA.
Email: berkr@sas.upenn.edu

**Keywords**

risk assessment, machine learning, fairness, criminal justice, discrimination

The use of actuarial risk assessments in criminal justice settings has of late been subject to intense scrutiny. There have been ongoing discussions about how much better in practice risk assessments derived from machine learning perform compared to risk assessments derived from older, conventional methods (Berk 2012; Berk and Bleich 2013; Brennan and Oliver 2013; Liu et al. 2011; Rhodes 2013; Ridgeway 2013a, 2013b). We have learned that when relationships between predictors and the response are complex, machine learning approaches can perform far better. When relationships between predictors and the response are simple, machine learning approaches will perform about the same as conventional procedures.

Far less close to resolution are concerns about fairness raised by the media (Angwin et al. 2016; Cohen 2012; Crawford 2016; Dieterich et al. 2016; Doleac and Stevenson 2016), government agencies (National Science and Technology Council 2016:30-32), foundations (Pew Center of the States 2011), and academics (Berk 2008; Berk and Hyatt 2015; Demuth 2003; Hamilton 2016; Harcourt 2007; Hyatt, Chanenson, and Bergstrom 2011; Starr 2014b; Tonry 2014).[1] Even when direct indicators of protected group membership, such as race and gender, are not included as predictors, associations between these measures and legitimate predictors can "bake in" unfairness. An offender's prior criminal record, for example, can carry forward earlier, unjust treatment not just by criminal justice actors but by an array of other social institutions that may foster disadvantage.

As risk assessment critic Sonja Starr (2014a) writes,

> While well intentioned, this approach [actuarial risk assessment] is misguided. The United States inarguably has a mass-incarceration crisis, but it is poor people and minorities who bear its brunt. Punishment profiling will exacerbate these disparities—including racial disparities—because the risk assessments include many race-correlated variables. Profiling sends the toxic message that the state considers certain groups of people dangerous based on their identity. It also confirms the widespread impression that the criminal justice system is rigged against the poor. (pp. A17)

On normative grounds, such concerns can be broadly legitimate, but without far more conceptual precision, it is difficult to reconcile competing claims and develop appropriate remedies. The debates can become rhetorical exercises, and few minds are changed.

This article builds on recent developments in computer science and statistics in which fitting procedures, often called algorithms, can assist criminal justice decision-making by addressing both accuracy *and* fairness.[2] Accuracy is formally defined by out-of-sample performance using one or more conceptions of prediction error (Hastie, Tibshirani, and Friedman 2009:section 7.2). There is no ambiguity. But, even when attempts are made to clarify what fairness can mean, there are several different kinds that can conflict with one another and with accuracy (Berk 2016b).

Examined here are different ways that fairness can be formally defined, how these different kinds of fairness can be incompatible, how risk assessment accuracy can be affected, and various algorithmic remedies that have been proposed. The perspectives represented are found primarily in statistics and computer science because those disciplines are the source of modern risk assessment tools used to inform criminal justice decisions.

No effort is made here to translate formal definitions of fairness into philosophical or jurisprudential notions in part because the authors of this article lack the expertise and in part because that multidisciplinary conversation is just beginning (Barocas and Selbst 2016; Ferguson 2015; Janssen and Kuk 2016; Kroll et al. 2017). Nevertheless, an overall conclusion will be that you can't have it all. Rhetoric to the contrary, challenging tradeoffs are required between different kinds of fairness and between fairness and accuracy.

Although for concreteness, criminal justice applications are the focus, the issues readily generalize to a very wide range of risk assessment applications. For example, decisions made by banks about whom to grant mortgage loans rest heavily on risk assessments for the chances that the loan will be repaid. Employers commonly do background checks to help determine whether a job applicant will be a reliable employee. Child welfare agencies typically decide when a minor should be placed in foster care based in part of the predicted risk from remaining in their current residence.

## Confusion Tables, Accuracy, and Fairness: A Prologue

For ease of exposition and with no important loss of generality, $Y$ is the response variable, henceforth assumed to be binary, and there are two legally protected group categories: men and women. We begin by introducing by example some key ideas needed later to define fairness and accuracy. We build on the simple structure of a $2 \times 2$ cross-tabulation (Berk 2016b; Chouldechova 2017; Hardt, Price, and Srebro 2016). Illustrations follow shortly.

**Table 1.** A Cross-tabulation of the Actual Outcome by the Predicted Outcome When the Prediction Algorithm Is Applied to a Data Set.

| Truth | Failure Predicted | Success Predicted | Conditional Procedure Error |
|---|---|---|---|
| Failure—a positive | $a$ | $b$ | $b/(a+b)$ |
| | True positives | False negatives | False negative rate |
| Success—a negative | $c$ | $d$ | $c/(c+d)$ |
| | False positives | True negatives | False positive rate |
| Conditional use error | $c/(a+c)$ | $b/(b+d)$ | $\frac{(c+b)}{(a+b+c+d)}$ |
| | Failure prediction error | Success prediction error | Overall procedure error |

Table 1 is a cross-tabulation of the actual binary outcome $Y$ by the predicted binary outcome $\hat{Y}$. Such tables are in machine learning often called a "confusion table" (also "confusion matrix"). $\hat{Y}$ is the fitted values that result when an algorithmic procedure is applied in the data. A "failure" is called a "positive" because it motivates the risk assessment; a positive might be an arrest for a violent crime. A "success" is a "negative," such as completing a probation sentence without any arrests. These designations are arbitrary but allow for a less abstract discussion.[3]

The left margin of the table shows the actual outcome classes. The top margin of the table shows the predicted outcome classes.[4] Cell counts internal to the table are denoted by letters. For example, "$a$" is the number of observations in the upper left cell. All counts in a particular cell have the same observed outcome class and the same predicted outcome class. For example, "$a$" is the number of observations for which the observed response class is a failure and the predicted response class is a failure. It is a true positive. Starting at the upper left cell and moving clockwise around the table are true positives, false negatives, true negatives, and false positives.

The cell counts and computed values on the margins of the table can be interpreted as descriptive statistics for the observed values and fitted values in the data on hand. Also common is to interpret the computed values on the margins of the table as estimates of the corresponding probabilities in a population. We turn to that later. For now, we just consider descriptive statistics.

There is a surprising amount of descriptive information that can be extracted from the table. We will use the following going forward.[5]

1. *Sample size*—The total number of observations conventionally denoted by $N$: $a + b + c + d$.
2. *Base rate*—The proportion of actual failures, which is $(a + b)/(a + b + c + d)$, or the proportion of actual successes, which is $(c + d)/(a + b + c + d)$.
3. *Prediction distribution*—The proportion predicted to fail and the proportion predicted to succeed: $(a + c)/(a + b + c + d)$ and $(b + d)/(a + b + c + d)$, respectively.
4. *Overall procedure error*—The proportion of cases misclassified: $(b + c)/(a + b + c + d)$.
5. *Conditional procedure error*—The proportion of cases incorrectly classified conditional on one of the two *actual* outcomes: $b/(a + b)$, which is the *false negative rate*, and $c/(c + d)$, which is the *false positive rate*.
6. *Conditional use error*—The proportion of cases incorrectly predicted conditional on one of the two *predicted* outcomes: $c/(a + c)$, which is the proportion of incorrect failure predictions, and $b/(b + d)$, which is the proportion of incorrect success predictions.[6] We use the term conditional *use* error because when risk is actually determined, the predicted outcome is employed; this is how risk assessments are used in the field.
7. *Cost ratio*—The ratio of false negatives to false positives $b/c$ or the ratio of false positives to false negatives $c/b$. When $b$ and $c$ are the same, the cost ratio is one, and false positives have same weight as false negatives. If $b$ is smaller than $c$, $b$ is *more* costly. For example, if $b = 20$ and $c = 60$, false negatives are three times more costly than false positives. One false negative is "worth" three false positives. In practice, $b$ can be more or less costly than $c$. It depends on the setting.

The discussion of fairness to follow uses all of these features of Table 1, although the particular features employed will vary with the kind of fairness. We will see, in addition, that the different kinds of fairness can be related to one another and to accuracy. But before getting into a more formal discussion, some common fairness issues will be illustrated with three hypothetical confusion tables.

Table 2 is a confusion table for a hypothetical set of women released on parole. Gender is the protected individual attribute. A failure on parole is a "positive," and a success on parole is a "negative." For ease of exposition, the counts are meant to produce a very simple set of results.

**Table 2.** Females: Fail or Succeed on Parole (Success Base Rate = 500/1,000 = .50, Cost Ratio = 200/200 = 1:1, and Predicted to Succeed 500/1,000 = .50).

| Truth | $\hat{Y}_{fail}$ | $\hat{Y}_{succeed}$ | Conditional Procedure Error |
|---|---|---|---|
| $Y_{fail}$—positive | 300<br>True positives | 200<br>False negatives | .40<br>False negative rate |
| $Y_{succeed}$—negative | 200<br>False positives | 300<br>True negatives | .40<br>False positive rate |
| Conditional use error | .40<br>Failure prediction error | .40<br>Success prediction error | |

The base rate for success is .50 because half of the women are not rearrested. The algorithm correctly predicts that the proportion who succeed on parole is .50. This is a favorable initial indication of the algorithm's performance because the marginal distribution of $Y$ and $\hat{Y}$ is the same.

Some call this "calibration" and assert that calibration is an essential feature of any risk assessment tool. Imagine the alternative: 70 percent of women on parole are arrest free, but the risk assessment projects that 50 percent will be arrest free. The instrument's credibility is immediately undermined. But calibration sets are a very high standard that existing practice commonly will fail to meet. Do the decisions of police officers, judges, magistrates, and parole boards perform at the calibration standard? Perhaps a more reasonable standard is that the any risk tool just needs perform better than current practice. Calibration in practice is different from calibration in theory, although the latter is a foundation for much formal work on risk assessment fairness. We will return to these issues later.[7]

The false negative rate and false positive rate of .40 are the same for successes and failures. When the outcome is known, the algorithm can correctly identify it 60 percent of the time. Usually, the false positive rate and the false negative rate are different, which complicates overall performance assessments.

Because here the number of false negatives and false positives is the same (i.e., 200), the cost ratio is 1 to 1. This too is empirically atypical. False negatives and false positives are equally costly according to the algorithm. Usually, they are not.

The prediction error of .40 is the same for predicted successes and predicted failures. When the outcome is predicted, the prediction is correct

**Table 3.** Males: Fail or Succeed on Parole (Success Base Rate = 500/1,500 = .33, Cost Ratio 400/200 = 2:1, and Predicted to Succeed 700/1,500 = .47).

| Truth | $\hat{Y}_{\text{fail}}$ | $\hat{Y}_{\text{succeed}}$ | Conditional Procedure Error |
|---|---|---|---|
| $Y_{\text{fail}}$—positive | 600<br>True positives | 400<br>False negatives | .40<br>False negative rate |
| $Y_{\text{succeed}}$—negatives | 200<br>False positives | 300<br>True negative | .40<br>False positive rate |
| Conditional use error | .25<br>Failure prediction error | .57<br>Success prediction error | |

60 percent of the time. Usually, prediction error will differ between predicted successes and predicted failures.

Each of these measures can play a role in fairness assessments. We do not consider fairness yet because Table 2 shows only the results for women. Fairness is addressed across two or more confusion tables one for each protected class.

Table 3 is a confusion table for a hypothetical set of men released on parole. To help illustrate fairness concerns, the base rate for success on parole is changed from .50 to .33. Men are substantially less likely to succeed on parole than women. The base rate was changed by multiplying the top row of cell counts in Table 2 by 2.0. That is the *only* change made to the cell counts. The bottom row of cell counts is unchanged.

Although the proportion of women predicted to succeed on parole corresponds to the actual proportion of women who succeed, the proportion of men predicted to succeed on is a substantial overestimate of the actual proportion of men who succeed. For men, the distribution of $Y$ is not the same as the distribution of $\hat{Y}$. There is a lack of calibration for men. Some might argue that this makes all the algorithmic results less defensible for men because an essential kind of accuracy has been sacrificed. (One would arrive at the same conclusion using predictions of failure on parole.) Fairness issues could arise in practice if decision makers, noting the disparity between the actual proportion who succeed on parole and the predicted proportion who succeed on parole, discount the predictions for men, implicitly introducing gender as an input to the decision to be made.

The false negative and false positive rates are the same and unchanged at .40. Just as for women, when the outcome is known, the algorithm can

correctly identify it 60 percent of the time. There are usually no fairness concerns when a confusion table measure being examined does not differ by protected class.

Failure prediction error is reduced from .40 to .25, and success prediction error is increased from .40 to .57. Men are more often predicted to succeed on parole when they actually do not. Women are more often predicted to fail on parole when they actually do not. If predictions of success on parole make a release more likely, some would argue that the prediction errors lead to decisions that unfairly favor men. Some would assert more generally that different prediction error proportions for men and women are by itself a source of unfairness.

Whereas in Table 2, .50 of the women are predicted to succeed overall, in Table 3, .47 of the men are predicted to succeed overall. This is a small disparity in practice, but it favors women. If decisions are affected, some would call this unfair, but it is a different source of unfairness than disparate prediction *errors* by gender.

Finally, although the cost ratio in Table 2 for women makes false positives and false negatives equally costly (1 to 1), in Table 3, false positives are twice as costly as false negatives. Incorrectly classifying a success on parole as failure is twice as costly for men (2 to 1). This too can be seen as unfair if it affects decisions. Put another way, individuals who succeed on parole but who would be predicted to fail are potentially of greater relative concern when the individual is a man.

It follows arithmetically that all of these potential unfairness and accuracy problems surface solely by changing the base rate even when the false negative rate and false positive rate are unaffected. Base rates can matter a great deal, a theme to which we will return. Base rates also matter substantially for a wide range of risk assessment settings such as those mentioned earlier. For example, diabetes base rates for Hispanics, blacks, and Native Americans can be as much as double the base rates for non-Hispanic whites (American Diabetes Association 2018). One consequence, other things equal, would be larger prediction errors for those groups when a diagnosis of diabetes is projected, implying a greater chance of false positives. The appropriateness of different medical interventions could be affected as a consequence.

We will see later that there are a number of proposals that try to correct for various kinds of unfairness, including those illustrated in the comparisons between Tables 2 and 3. For example, it is sometimes possible to tune classification procedures to reduce or even eliminate some forms of unfairness.

**Table 4.** Males Tuned: Fail or Succeed on Parole (Success Base Rate = 500/1,500 = .33, Cost Ratio = 200/200 = 1:1, and Predicted to Succeed 500/1,500 = .33).

| Truth | $\hat{Y}_{\text{fail}}$ | $\hat{Y}_{\text{succeed}}$ | Conditional Procedure Error |
|---|---|---|---|
| $Y_{\text{fail}}$—positive | 800<br>True positives | 200<br>False negatives | .20<br>False negative rate |
| $Y_{\text{succeed}}$—negative | 200<br>False positives | 300<br>True negatives | .40<br>False positive rate |
| Conditional use error | .20<br>Failure prediction error | .40<br>Success prediction error | |

In Table 4, for example, the success base rate for men is still .33, but the cost ratio for men is tuned to be 1 to 1. Now, when success on parole is predicted, it is incorrect 40 times of the 100 and corresponds to .40 success prediction error for women. When predicting success on parole, one has equal accuracy for men and women. A kind of unfairness has been eliminated. Moreover, the fraction of men predicted to succeed on parole now equals the actual fraction of men who succeed on parole. There is calibration for men. Some measure of credibility has been restored to their predictions.

However, the false negative rate for men is now .20, not .40, as it is for women. In trade, therefore, when men actually fail on parole, the algorithm is more likely than for women to correctly identify it. By this measure, the algorithm performs better for men. Trade-offs like these are endemic in classification procedures that try to correct for unfairness. Some trade-offs are inevitable, and some are simply common. This too is a theme to which we will return.

## The Statistical Framework

We have considered confusion tables as descriptive tools for data on hand. The calculations on the margins of the table are proportions. Yet those proportions are often interpreted as probabilities. Implicit are properties that cannot be deduced from the data alone. Commonly, reference to a data generation process is required (Berk 2016a:section 1.4; Kleinberg, Mullainathan, and Raghavan 2016). For clarity and completeness, we need to consider that data generation process.

There are practical concerns as well requiring a "generative" formulation. In many situations, one wants to draw inferences beyond the data being

**Table 5.** Notation for Statistical Concepts.

| Notation | Meaning |
|---|---|
| $Y$ | Response variable (numeric or categorical) |
| $L$ | Legitimate predictors |
| $S$ | Predictors for protected classes |
| $P(Y, L, S)$ | Parent joint probability distribution |
| $E(Y\|L, S)$ | True response surface |
| $f(F, S)$ | True response function |
| $f^*(L, S)$ | True response function approximation |
| $h(L, S)$ | Fitting procedure |
| $\hat{f}^*(L, S)$ | Fitted approximation structure |
| $\hat{Y}$ | Fitted approximate values of the response |

analyzed. Then, the proportions can be seen as statistical estimates. For example, a confusion table for release decisions at arraignments from a given month might be used to draw inferences about a full year of arraignments in that jurisdiction (Berk and Sorenson 2016). Likewise, a confusion table of the housing decisions made for prison inmates (e.g., low-security housing vs. high-security housing) from a given prison in a particular jurisdiction might be used to draw inferences about placement decisions in other prisons in the same jurisdiction (Berk and de Leeuw 1999).[8] But perhaps most important, algorithmic results from a given data set are commonly used to inform decisions in the future. Generalizations are needed over time.

Under such circumstances, one needs a formal rationale for how the data came to be and for the estimation target. In conventional survey sample terms, one must specify a population and one or more population parameters whose values are to be estimated from the data. Probability sampling then provides the requisite justification for statistical inference.

There is a broader formulation that is usually more appropriate for algorithmic procedures. The formulation has each observation randomly realized from a single joint probability distribution. This is a common approach in computer science, especially for machine learning (Bishop 2006: section 1.5; M. J. Kearns 1994:section 1.2), and also can be found in econometrics (White 1980) and statistics (Buja et al. 2018a; Freedman 1981).

Table 5 summarizes the notation to follow. We denote the parent, joint probability distribution by $P(Y, L, S)$. $Y$ is the outcome of interest. An arrest while on probation is an illustration. $L$ includes "legitimate" predictors such as prior convictions. $S$ includes "protected" predictors such as race, ethnicity, and gender, sometimes called "suspect" variables.[9] In computer science,

$P(Y, L, S)$ often is called a "target population." The data on hand are a set of IID realized observations from $P(Y, L, S)$.[10] In some branches of computer science, such as machine learning, each realized observation is called an "example."[11]

$P(Y, L, S)$ has all of the usual moments, which is a harmless assumption in practice. From this, the population can be viewed as the limitless number of observations that could be realized from the joint probability distribution (as if by random sampling), each observation an IID realized case. Under this conception of a population, all moments and conditional moments are necessarily expectations.

Because $P(Y, L, S)$ is the source of the data, features of this joint probability distribution shape the estimation enterprise. There is in the target population some true function of $L$ and $S$, $f(L, S)$, linking the predictors to the conditional expectations of $Y$: $E(Y|L, S)$. When $Y$ is categorical, these conditional expectations are conditional probabilities. $E(Y|L, S)$ is the "true response surface."

The true response function is assumed to be unknown.[12] However, there is in the population an *approximation*, $f^*(L, S)$, of the true response function. It can be called the "best" approximation if in principle it could be a product of some loss minimization procedure such as appropriate estimating equations (Buja et al. 2018b).[13] The approximation is then "best" by that criterion.

The best approximation $f^*(L, S)$ is specified pragmatically from the data on hand. In a modeling setting, it might be called a "working model." In an algorithmic setting, it might be called a "heuristic." No claims are made that it corresponds to the true response function.[14]

A fitting procedure, $h(L, S)$, is applied to the data. The result $\hat{f}^*(L, S)$ is an estimate of $f^*(L, S)$, not $f(L, S)$. Often, one is able to compute asymptotically unbiased estimates of key $f^*(L, S)$ features (e.g., generalization error) coupled with valid statistical tests and confidence intervals (Buja et al. 2018b).[15] Usually, the most important feature of $f^*(L, S)$ estimated by $\hat{f}^*(L, S)$ is the fitted values $\hat{Y}$. They are estimates of the true response surface *approximation*. It is often possible to compute point-by-point confidence intervals for each $\hat{Y}$, understanding that these are proper intervals for the approximate response surface, not the true response surface.

Consider a simple illustration to help fix these ideas. Suppose the true response function in some population is a high-order polynomial conditioned on several predictors. The true response surface lies in the space defined by those predictors and is comprised of the expectation of $Y$ for each configuration of $x$ values. There is a population approximation of the true response function computed in principle by least squares that is a linear function of the

same predictors. The linear function is then the best linear approximation of the true response function. Its fitted values, for each configuration of $x$ values, constitute an approximation in the population of the true response surface. Data are realized from the joint probability distribution as if by simple random sampling. A working linear regression model is specified using the same predictors and the same response. Its regression coefficients and disturbance variance are estimated by way of least squares. The regression fitted values for each configuration of $x$ values are estimates of the approximation response surface. All of these estimates are asymptotically unbiased for the approximation. Valid statistical tests and confidence intervals can follow. Perhaps the major difference in practice is that far more flexible fitting procedures than linear regression are employed, and the true response surface can be far more complex than a polynomial.

Whether such conceptual scaffolding makes sense for real data depends on substantive knowledge and knowledge about how the data were actually produced. For example, one might be able to make the case that for a particular jurisdiction, all felons convicted in a given year can usefully be seen as IID realizations from the population of all convicted felons that could have been produced that year and perhaps for a few years before and a few years after. One would need to argue that for the given year, or proximate years, there were no meaningful changes in any governing statutes, the composition of the sitting judges, the mix of felons, and the practices of police, prosecutors, and defense attorneys. A more detailed consideration would for this article be a distraction and is discussed elsewhere in an accessible linear regression setting (Berk, Brown et al. 2017).

## Defining Fairness

### *Definitions of Algorithmic Fairness*

We are now ready to consider definitions of algorithmic fairness. Instructive definitions can be found in computer science (Calmon et al. 2017; Chouldechova 2017; Dwork et al. 2012; Friedler, Scheidegger, and Venkatasubramanian 2016; Hardt et al. 2016; Joseph et al. 2016; Kamishima, Akaho, and Sakuma 2011; Kamiran and Calders 2012; Kleinberg et al. 2016; Pedreschi, Ruggieri, and Turini 2008), criminology (Angwin et al. 2016; Berk 2016b; Dieterich, Mendoza, and Brennan 2016), and statistics (Corbett-Davies et al. 2017; Johnson, Foster, and Stine 2016; Johndrow and Lum 2017). All are recent and focused on algorithms used to inform real-world decisions in both public and private organizations.

Each definition is broadly similar in intent. What matters is some definition of equality for protected groups. But the definitions can differ in substantive and technical details. There can be frustrating variation in notation combined with subtle differences in how key concepts are operationalized. There can also be a conflation of training data properties, performance of an algorithm, and decisions that can follow.[16] We focus here on algorithms and the data on which they are trained. How actions are affected is a very important, but different, matter (Berk 2017; Kleinberg et al. 2017).

Much of the formal theory on algorithmic fairness is derived for risk instruments that output some form of risk score. Often, the risk score is a probability. An outcome class is assigned by imposing some threshold on the risk score. For example, if for a given offender, the instrument's probability of an arrest on parole is greater than .50, a class of "high risk" can be assigned. Our discussion of fairness is agnostic about how outcome classes are assigned by an algorithm and about the algorithmic procedure used. This makes the discussion more general and, we hope, more accessible. As before, we proceed with confusion tables.

In order to provide clear definitions of algorithmic fairness, we will proceed for now as if $\hat{f}^{*}(L,S)$ estimates are the same as the corresponding population values. In this way, we do not complicate a discussion of fairness with concerns about *estimation* (i.e., inferences from the realized data to features of the joint probability distribution). Estimation is addressed later. The notation is drawn from Table 1, but there will be a separate confusion table for each class in the protected group. Comparisons are made between these tables. Consistent with much of the extant fairness literature, we build on algorithmic accuracy rather than algorithmic error.[17] As before, a failure on parole is called a positive and a success on parole is called a negative.

1. *Overall accuracy equality* is achieved by $\hat{f}^{*}(L,S)$ when overall procedure accuracy is the same for each class of a protected group (e.g., men and women). That is, $(a+d)/(a+b+c+d)$ should be the same (Berk 2016b). This definition assumes that true negatives are as desirable as true positives. In many settings they are not, and a cost-weighted approach is required. For example, true negatives (i.e., successes on parole) may be twice as desirable as true positives (i.e., failures on parole). Or put another way, false negatives may be two times less desirable than false positives. Overall accuracy equality is not commonly used because it does not distinguish between accuracy for positives and accuracy for negatives. Nevertheless, it has been mentioned in some media accounts (Angwin et al. 2016) and is

related in spirit to "accuracy equity" as used by Dieterich and colleagues (2016).[18]

2. *Statistical parity* is achieved by $\hat{f}^{*}(L,S)$ when the marginal distributions of the predicted outcome classes are the same for each class of a protected group (e.g., Muslims and Christians). That is, $(a+c)/(a+b+c+d)$ and $(b+d)/(a+b+c+d)$, although typically different from one another, are the same for both protected group classes (Berk 2016b). For example, the proportion of inmates predicted to succeed on parole (i.e., negatives) should be the same for Muslim and Christian parolees. When this holds, it also holds for predictions of failure on parole (i.e., positives) because the outcome is binary. This definition of statistical parity, sometimes called "demographic parity," has been criticized because it can lead to highly undesirable decisions for individuals (Dwork et al. 2012). One might incarcerate Muslims who pose no public safety risk so that the same proportions of Muslims and Christians are released on parole. Our definition is much like statistical parity as defined by Chouldechova (2017:4), although she requires an underlying risk score with the "high-risk" class determined by score values above a certain threshold.
3. *Conditional procedure accuracy equality* is achieved by $\hat{f}^{*}(L,S)$ when conditional procedure accuracy is the same for both protected group classes (Berk 2016b). In our notation, $a/(a+b)$ is the same for, say, African Americans and whites, and $d/(c+d)$ is the same for African Americans and whites. Conditioning on the known outcome, is $\hat{f}^{*}(L,S)$ equally accurate across protected group classes? This is the same as considering whether the false negative rate and the false positive rate, respectively, are the same for African Americans and whites. Conditional procedure accuracy equality is a common concern in criminal justice applications (Dieterich et al. 2016). Hardt and his colleagues (2016:2-3) use the term "equalized odds" for a closely related definition, and there is a special case they call "equality of opportunity" that effectively is the same as our conditional procedure accuracy equality, but only for the outcome class that is more desirable.[19] Chouldechova (2017:4) uses the term "error rate balance" for conditional procedure accuracy equality, but, as before, requires a threshold on a risk score to arrive at a high-risk class.
4. *Conditional use accuracy equality* is achieved by $\hat{f}^{*}(L,S)$ when conditional use accuracy is the same for both protected group classes (Berk 2016b). One is conditioning on the algorithm's *predicted*

outcome not the actual outcome. That is, $a/(a+c)$ is the same for individuals, say, born in the United States and for U.S. citizens born elsewhere, and $d/(b+d)$ is the same for individuals born in the United States and for U.S. citizens born elsewhere. Conditional use accuracy equality has also been a common concern in criminal justice risk assessments (Dieterich et al. 2016). Conditional on the prediction of success (or failure), is the projected probability of success (or failure) the same across protected group classes? If not, membership in a protected groups class is associated with conditional use accuracy. Chouldechova (2017:3-4) calls this "predictive parity." Chouldechova would also say that in our confusion table setting, a risk instrument delivering predictive parity is "well calibrated."[20]

5. *Treatment equality* is achieved by $\hat{f}^{*}(L,S)$ when the ratio of false negatives and false positives (i.e., $c/b$ or $b/c$) is the same for both protected group categories. The term "treatment" is used to convey that such ratios can be a policy lever with which to achieve other kinds of fairness. For example, if false negatives are treated as more costly for men than women so that conditional procedure accuracy equality can be achieved, men and women are being treated differently by the algorithm (Feldman et al. 2015). Incorrectly classifying a failure on parole as a success (i.e., a false negative), say, is a bigger mistake for men. The relative numbers of false negatives and false positives across protected group categories also can by itself be viewed as a problem in criminal justice risk assessments (Angwin et al. 2016). Chouldechova (2017:4) addresses similar issues, but through the false negative rate and the false positive rate: our $b/(a+b)$ and $c/(c+d)$, respectively.
6. *Total fairness* is achieved by $\hat{f}^{*}(L,S)$ when (1) overall accuracy equality, (2) statistical parity, (3) conditional procedure accuracy equality, (4) conditional use accuracy equality, and (5) treatment equality are all achieved. Although a difficult pill for some to swallow, we will see that in practice, total fairness cannot be achieved except in highly artificial simulations.

Each of the definitions of fairness applies when there are more than two outcome categories. However, there are more statistical summaries that need to be reviewed. For example, when there are three response classes, there are three ratios of false negatives to false positives to be examined.

There are also other definitions of fairness not discussed because they currently cannot be operationalized in a useful manner. For example, nearest

neighbor parity is achieved if similarly situated individuals are treated similarly (Dwork et al. 2012). Similarly situated is measured by the Euclidian distance between the individuals in predictor space. Unfortunately, the units in which the predictors are measured can make an important difference, and standardizing them just papers over the problem. Nevertheless, the ideas introduced are very important from both a statistical and jurisprudential point of view.[21]

## Estimation Accuracy

We build again on work by Buja and his colleagues (2018a). When the procedure $h(L, S)$ is applied to the IID data, some will argue that the estimation target is the true response surface. But even asymptotically, there is no credible claim that the true response surface is being estimated in an unbiased manner. The same applies to the probabilities from a cross-tabulation of $Y$ by $\hat{Y}$.

With larger samples, the random estimation error is smaller. On the average, the estimates are closer to the truth. However, the gap between the estimates and the truth combines bias and variance. That gap is not a conventional confidence interval, nor can it be transformed into one. One would have to remove the bias, and to remove the bias, one would need to compare the estimates to the unknown truth.

Alternatively, the estimation target for $h(L, S)$ can be an acknowledged approximation of the true response surface. In the population, the approximation has the same structure as $h(L, S)$ so that the algorithm becomes a plug-in estimator. Therefore, estimates of probabilities from Table 1 can be estimates of the corresponding probabilities from a $Y$ by $\hat{Y}$ cross-tabulation as if $h(L, S)$ were applied in the population. Thanks to the IID nature of the data, these estimates can also be asymptotically unbiased so that in large samples, the bias will likely be relatively small. This allows one to use sample results to address fairness as long as one appreciates that it is fairness measured by the approximation, not the true $E(Y|L, S)$.

Estimation accuracy is addressed by out-sample performance. Fitted values in-sample will be subject to overfitting. In practice, this means using test data, or an approximation thereof (e.g., cross-validation), with measures of fit such as generalization error or expected prediction error (Hastie et al. 2009: section 7.2). Often, good estimates of accuracy may be obtained, but the issues can be tricky. Depending on the procedure $h(L, S)$ and the availability of an instructive form of test data, there are different tools that vary in their assumptions and feasibility (Berk 2016a). With our focus on fairness, such details are a diversion.[22] The fairness issues are unchanged.

## Trade-offs

We turn to trade-offs and begin by emphasizing an obvious point that can get lost in discussions of fairness. If the goal of applying $h(L, S)$ is to capitalize on nonredundant associations that $L$ and $S$ have with the outcome, excluding $S$ will reduce accuracy. Any procedure that even just discounts the role of $S$ will lead to less accuracy. The result is a larger number of false negatives and false positives. For example, if $h(L)$ is meant to help inform parole release decisions, there will likely be an increase in both the number of inmates who are unnecessarily detained and the number of inmates who are inappropriately released. The former victimizes inmates and their families. The latter increases the number of crime victims. But fairness counts too, so we need to examine trade-offs.

Because the different kinds of fairness defined earlier share cell counts from the cross-tabulation of $Y$ against $\hat{Y}$, and because there are relationships between the cell counts themselves (e.g., they sum to the total number of observations), the different kinds of fairness are related as well. It should not be surprising, therefore, that there can be trade-offs between the different kinds of fairness. Arguably, the trade-off that has gotten the most attention is between conditional use accuracy equality and the false positive and false negative rates (Angwin et al. 2016; Chouldechova 2017; Dieterich et al. 2016; Kleinberg et al. 2016; Pleiss et al. 2017). It is also the trade-off that to date has the most complete mathematical results.

### *Some Proven "Impossibility Theorems"*

We have conveyed informally that there are incompatibilities between different kinds of fairness. It is now time to be specific. We begin with three definitions. They will be phrased in probability terms but are effectively the same if phrased in terms of proportions.

- *Calibration*—Calibration was introduced earlier. Suppose an algorithm produces a probability risk score that can then be used to assign an outcome class. "Calibration within groups requires that for each group $t$, and each bin $b$ with associated score $\upsilon_b$, the expected number of people from group $t$ in $b$ who belong to the positive class should be a $\upsilon_b$ fraction of the expected number of people from group $t$ assigned to $b$ (Kleinberg et al. 2016:4). Here, a "positive" would be a rearrest on parole. For example, if the risk score in a probability metric is .37, the predicted probability of recidivism should also be .37.

Calibration can become a fairness matter if there is calibration within one group but not within the other. A decision maker may be inclined to take the predictions less seriously for the group that lacks calibration (Kleinberg et al. 2016; Pleiss et al. 2017). Essentially, the same fairness issues arise under Chouldechova's (2017) definition of predictive parity, although the risk score does not have to be in a probability metric, and for our definition of conditional use accuracy equality, within a confusion table formulation.

- *Base rate*—This too was introduced earlier. In the population, base rates are determined by the marginal distribution of the response, defined by the probability of each outcome class. For example, the base rate for succeeding on parole might be .65 and for not succeeding on parole is then .35. If there are $C$ outcome classes, there will be $C$ base rate probabilities.

We are concerned here with base rates for different protected group classes, such as men compared to women. Base rates for each protected group class are said to be equal if they are identical. Is the probability of succeeding on parole .65 for both men and women?

- *Separation*—In a population, the observations are separable if for each possible configuration of predictor values, there is some $h(L, S)$ for which the probability of membership in a given outcome class is always either 1.0 or 0.0. In other words, perfectly accurate classification is possible. In practice, what matters is whether there is perfect classification when $h(L, S)$ is applied to data.

And now the impossibility results: When the base rates differ by protected group and when there is not separation, one cannot have both calibration and equality in the false negative and false positive rates (Kleinberg et al. 2016). For Chouldechova's formulation (2017:5), if "the base rate differs across groups, any instrument that satisfies predictive parity at a given threshold . . . *must* have imbalanced false positive or false negative errors rates at that threshold" (emphasis in the original). Within our formulation, when base rates for protected group classes differ, one cannot have simultaneously conditional use accuracy equality and, across protected group classes, equal false positive and false rates.

The implications of the impossibility results are huge. First, if there is variation in base rates and no separation, you can't have it all. The goal of complete race or gender neutrality is unachievable. In practice, both requirements are virtually never met, except in highly stylized examples.

**Table 6.** Males: A Cross-tabulation When All Cases Are Assigned the Outcome of Failure (Base Rate = .80, $N = 500$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 400 | 0 | 1.0 |
| $Y = 0$ (a negative—not fail) | 100 | 0 | 0.0 |
| Conditional use accuracy | .80 | — | |

Second, altering a risk algorithm to improve matters can lead to difficult stakeholder choices. For example, if it is essential to have conditional use accuracy equality, the algorithm will produce different false positive and false negative rates across the protected group categories. Conversely, if it is essential to have the same rates of false positives and false negatives across protected group categories, the algorithm cannot produce conditional use accuracy equality. Stakeholders will have to settle for an increase in one for a decrease in the other.

Third, there are other kinds of fairness that are in play but are unaddressed in the existing proofs. To arrive at mathematically tractable problems, simplifications are often required. We will see shortly that this affects the algorithmic remedies proposed and their potential reception by stakeholders. For example, anything that downweights the importance of certain predictors for one protected group but not another can introduce a form of treatment inequality. In service of statistical parity, for instance, one might give less weight to prior arrests for men than for women even though the consequences of prior crimes for crime victims are the same and conditional use accuracy likely will be reduced for men. To see how these moving parts can interact, consider the following didactic illustrations.

*Trivial case #1: Assigning the same outcome class to all.* Suppose $h(L, S)$ assigns the same outcome class to everyone (e.g., a failure). Such an assignment procedure would never be used in practice, but it raises some important issues in a simple setting. Tables 6 and 7 provide an example when the base rates are the same for men and women. There are 500 men and 50 women, but the relative representation of men and women does not matter materially in what follows. Failures are coded 1 and successes are coded 0, much as they might be in practice. Each case is assigned a failure (i.e., $\hat{Y} = 1$), but the same lessons would be learned if each case is assigned a success (i.e., $\hat{Y} = 0$). A base rate of .80 for failures is imposed on both tables.

In practice, this approach makes no sense. Predictors are not being exploited. But, one can see that there is conditional procedure accuracy

**Table 7.** Females: A Cross-tabulation When All Cases Are Assigned the Outcome of Failure (Base Rate = .80, $N = 50$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 40 | 0 | 1.0 |
| $Y = 0$ (a negative—not fail) | 10 | 0 | 0.0 |
| Conditional use accuracy | .80 | — | |

**Table 8.** Males: A Cross-tabulation With Failure Assigned to All With a Probability of .30 (Base Rate = .80, $N = 500$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 120 | 280 | .30 |
| $Y = 0$ (a negative—not fail) | 30 | 70 | .70 |
| Conditional use accuracy | .80 | .20 | |

equality, conditional use accuracy equality, and overall accuracy equality. The false negative and false positive rates are the same for men and women as well at 0.0 and 1.0. There is also statistical parity. One does very well on fairness for a risk tool that cannot help decision-makers address risk in a useful manner. Accuracy has been sacrificed in service of fairness. The dramatic trade-off between accuracy and fairness has come down definitively on the side of fairness.

If one allows the base rates for men and women differ, there is immediately a fairness price. Suppose in Table 6, 500 men fail instead of 400. The false positive and false negative rates are unchanged. But because the base rate for men is now larger than the base rate for women (i.e., .83 vs. .80), conditional use accuracy is now higher for men (i.e., also .83), and a lower proportion of men will be incorrectly predicted to fail (i.e., .17). This is the sort of result that would likely trigger charges of gender bias. Even in this "trivial" case, base rates matter.[23]

*Trivial case #2: Assigning the classes using the same probability for all.* Suppose each case is assigned to an outcome class with the *same* probability. As in trivial case #1, no use is made of predictors, so that accuracy does not figure into the fitting process.

For Tables 8 and 9, the assignment probability for failure is .30 for all, and therefore, the assignment probability for success is .70 for all. Nothing

**Table 9.** Females: A Cross-tabulation With Failure Assigned to All With a Probability of .30 (Base Rate = .80, $N = 50$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 12 | 28 | .30 |
| $Y = 0$ (a negative—not fail) | 3 | 7 | .70 |
| Conditional use accuracy | .80 | .20 | |

**Table 10.** Males: A Cross-tabulation With Separation and Perfect Prediction (Base Rate = .80, $N = 500$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 400 | 0 | 1.0 |
| $Y = 0$ (a negative—not fail) | 0 | 100 | 1.0 |
| Conditional use accuracy | 1.0 | 1.0 | |

important changes should some other probability be used.[24] The base rates for men and women are the same. For both, the proportions that fail are .80.

In Tables 8 and 9, we have the same fairness results we had in Tables 6 and 7, again with accuracy sacrificed. But suppose the second row of entries in Table 9 were 30 and 70 rather than 3 and 7. Now, the failure base rate for women is .29, not .80. Conditional procedure accuracy equality remains from which it follows that the false negative and false positive rates are the same as well. But conditional use accuracy equality is lost. The probabilities of correct predictions for men are again .80 for failures and .20 for successes. But for women, the corresponding probabilities are .29 and .71. Base rates can really matter.

*Perfect separation.* We now turn to an $h(L, S)$ that is not trivial, but also very unlikely in practice. In a population, the observations are separable. In Tables 10 and 11, there is perfect separation, and $h(L, S)$ finds it. Base rates are the same for men and women: .80 fail.

There are no false positives or false negatives, so the false positive rate and the false negative rate for both men and women are 0.0. There is conditional procedure accuracy equality and conditional use accuracy equality because both conditional procedure accuracy and conditional use accuracy are perfect. This is the ideal, but fanciful, setting in which we can have it all.

Suppose for women in Table 11, there are 20 women who do not fail rather than 10. The failure base rate for females is now .67 rather than

**Table 11.** Females: A Cross-tabulation With Separation and Perfect Prediction (Base Rate = .80, $N = 50$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 40 | 0 | 1.0 |
| $Y = 0$ (a negative—not fail) | 0 | 10 | 1.0 |
| Conditional use accuracy | 1.0 | 1.0 | |

**Table 12.** Females: A Cross-tabulation Without Separation (Base Rate = .56, $N = 900$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 300 | 200 | .60 |
| $Y = 0$ (a negative—not fail) | 200 | 200 | .50 |
| Conditional use accuracy | .60 | .50 | |

**Table 13.** Males: Confusion Table Without Separation (Base Rate = .56, $N = 1{,}400$).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 600 | 400 | .60 |
| $Y = 0$ (a negative—not fail) | 400 | 400 | .50 |
| Conditional use accuracy | .60 | .50 | |

.80. But because of separation, conditional procedure accuracy equality and conditional use accuracy equality remain, and the false positive and false negative rates for men and women are still 0.0. Separation saves the day.[25]

*Closer to real life.* There will virtually never be separation in the real data even if there there happens to be separation in the joint probability distribution responsible for the data. The fitting procedure $h(L, S)$ may be overmatched because important predictors are not available or because the algorithm arrives at a suboptimal result. Nevertheless, some types of fairness can sometimes be achieved if base rates are cooperative.

If the base rates are the same and $h(L, S, )$ finds that, there can be lots of good news. Tables 12 and 13 illustrate. Conditional procedure accuracy equality, conditional use accuracy equality, and overall procedure accuracy hold, and the false negative rate and the false positive rate are the same for

**Table 14.** Confusion Table for Females With No Separation and a Different Base Rate Compared to Males (Female Base Rate Is 500/900 = .56).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 300 | 200 | .60 |
| $Y = 0$ (a negative—not fail) | 200 | 200 | .50 |
| Conditional use accuracy | .60 | .50 | |

**Table 15.** Confusion Table for Males With No Separation and a Different Base Rate Compared to Females (Male Base Rate Is 1,000/2,200 = .45).

| Truth | $\hat{Y} = 1$ | $\hat{Y} = 0$ | Conditional Procedure Accuracy |
|---|---|---|---|
| $Y = 1$ (a positive—fail) | 600 | 400 | .60 |
| $Y = 0$ (a negative—not fail) | 600 | 600 | .50 |
| Conditional use accuracy | .50 | .40 | |

men and women. Results like those shown in Tables 12 and 13 can occur in real data but would be rare in criminal justice applications for the common protected groups. Base rates will not be the same.

Suppose there is separation, but the base rates are not the same. We are back to Tables 10 and 11, but with a lower base rate. Suppose there is no separation, but the base rates are the same. We are back to Tables 12 and 13.

From Tables 14 and 15, one can see that when there is no separation and different base rates, there can still be conditional procedure accuracy equality. From conditional procedure accuracy equality, the false negative rate and false positive rate, though different from one another, are the same across men and women. This is a start. But treatment equality is gone from which it follows that conditional use accuracy equality has been sacrificed. There is greater conditional use accuracy for women.

Of the lessons that can be taken from the sets of tables just analyzed, perhaps the most important for policy is that when there is a lack of separation and different base rates across protected group categories, a key tradeoff will exist between the false positive and false negative rates on one hand and conditional use accuracy equality on the other. Different base rates across protected group categories would seem to require a thumb on the scale if conditional use accuracy equality is to be achieved. To see if this is true, we now consider corrections that have been proposed to improve algorithmic fairness.

## Potential Solutions

There are several recent papers that have proposed ways to reduce and even eliminate certain kinds of bias. As a first approximation, there are three different strategies (Hajian and Domingo-Ferrer 2013), although they can also be combined when accuracy as well as fairness are considered.

### *Preprocessing*

Preprocessing means eliminating any sources of unfairness in the data before $h(L, S)$ is formulated. In particular, there can be legitimate predictors that are related to the classes of a protected group. Those problematic associations can be carried forward by the algorithm.

One approach is to remove all linear dependence between $L$ and $S$ (Berk 2008). One can regress in turn each predictor in $L$ on the predictors in $S$ and then work with the residuals. For example, one can regress predictors such as prior record and current charges on race and gender. From the fitted values, one can construct "residualized" transformations of the predictors to be used.

A major problem with this approach is that interactions effects (e.g., with race and gender) containing information leading to unfairness are not removed unless they are explicitly included in the residualizing regression *even if all of the additive contaminants are removed*. In short, all interactions effects, even higher order ones, would need to be anticipated. The approach becomes very challenging if interaction effects are anticipated between $L$ and $S$.

Johndrow and Lum (2017) suggest a far more sophisticated residualizing process. Fair prediction is defined as constructing fitted values for some outcome using no information from membership in any protected classes. The goal is to transform all predictors so that fair prediction can be obtained "while still preserving as much 'information' in X as possible" (Johndrow and Lum 2017:6). They formulate this using the Euclidian distance between the original predictors and the transformed predictors. The predictors are placed in order of the complexity of their marginal distribution, and each is residualized in turn using as predictors results from previous residualizations and indicators for the protected class. The regressions responsible for the residualizations are designed to be flexible so that nonlinear relationships can be exploited. However, interaction variables can be missed. For example, race can be removed from gang membership and from age, but not necessarily their product—being young *and* a gang member can still be associated with race. Also, as Johnson and Lum note, they are only able to consider one

form of unfairness. Consequently, they risk exacerbating one form of unfairness while mitigating another.

Base rates that vary over protected group categories can be another source of unfairness. A simple fix is to rebalance the marginal distributions of the response variable so that the base rates for each category are the same. One method is to apply weights for each group separately so that the base rates across categories are the same. For example, women who failed on parole might be given more weight and males who failed on parole might be given less weight. After the weighting, men and women could have a base rate that was the same as the overall base rate.

A second rebalancing method is to randomly relabel some response values to make the base rates comparable. For example, one could for a random sample of men who failed on parole, recode the response to a success and for a random sample of women who succeeded on parole, recode the response to a failure.

Rebalancing has at least two problems. First, there is likely to be a loss in accuracy. Perhaps such a trade-off between fairness and accuracy will be acceptable to stakeholders, but before such a decision is made, the trade-off must be made numerically specific. How many more armed robberies, for instance, will go unanticipated in trade for a specified reduction in the disparity between incarceration rates for men and women? Second, rebalancing implies using different false positive to false negative rates for different protected group categories. For example, false positives (e.g., incorrectly predicting that individuals will fail on parole) are treated as relatively more serious errors for men than for women. In addition to the loss in accuracy, stakeholders are trading one kind of unfairness for another.

A third approach capitalizes on association rules, popular in marketing studies (Hastie et al. 2009:section 14.2). Direct discrimination is addressed when features of some protected class are used as predictors (e.g., male). Indirect discrimination is addressed when predictors are used that are related to those protected classes (e.g., prior arrests for aggravated assault). There can be evidence of either if the conditional probability of the outcome changes when either direct or indirect measures of protected class membership are used as predictors compared to when they are not used. One potential correction can be obtained by perturbing the suspect class membership (Pedreschi et al. 2008). For a random set of cases, one might change the label for men to the label for a woman. Another potential correction can be obtained by perturbing the outcome label. For a random set of men, one might change failure on parole to success on parole (Hajian and Domingo-Ferrer 2013). Note that the second approach changes the base rate. We examined earlier

the consequences of changing base rates. Several different kinds of fairness can be affected. It can be risky to focus on a single definition of fairness.

A fourth approach is perhaps the most ambitious. The goal is to randomly transform all predictors except for indicators of protected class membership so that the joint distribution of the predictors is less dependent on protected class membership. An appropriate reduction in dependence is a policy decision. The reduction of dependence is subject to two constraints: (1) the joint distribution of the transformed variables is very close to the joint distribution of the original predictors, and (2) no individual cases are substantially distorted because large changes are made in predictor values (Calmon et al. 2017). An example of a distorted case would be a felon with no prior arrests assigned a predictor value of 20 prior arrests. It is unclear, however, how this procedure maps to different kinds of fairness. For example, the transformation itself may inadvertently treat prior crimes committed by men as less serious than similar prior crimes committed by women—the transformation may be introducing the prospect of unequal treatment. There are also concerns about the accuracy price, which is not explicitly taken into account. Finally, there is no allowance for interaction effects related to protected class membership unless all of the relevant product variables are included in the set of predictors. And even if such knowledge were available, the number of columns in the matrix of predictors could become enormous, and very high levels of multicollinearity would follow.

### *In-processing*

In-processing means building fairness adjustments into $h(L, S)$. To take a simple example, risk forecasts for particular individuals that have substantial uncertainty can be altered to improve fairness. If whether or not an individual is projected as high risk depends on little more than a coin flip, the forecast of high risk can be changed to low risk to serve some fairness goal. One might even order cases from low certainty to high certainty for the class assigned so that low certainty observations are candidates for alterations first. The reduction in out-of-sample accuracy may well be very small. One can embed this idea in a classification procedure so that explicit trade-offs are made (Corbett-Davies et al. 2017; Kamiran and Calders 2009, 2012). But this too can have unacceptable consequences for the false positive and false negative rates. A thumb is being put on the scale once again. There is inequality of treatment.

An alternative approach is to add a new penalty term to a penalized fitting procedure. Kamishima and colleagues (2011) introduce a fairness regularizer

into a logistic regression formulation that can penalize the fit for inappropriate associations between membership in a protected group class and the response or legitimate predictors. However, this too can easily lead to unequal treatment.

Rather than imposing a fairness penalty, one can impose fairness constraints. Agarwal and colleagues (2018) define a "reduction" that treats the accuracy-fairness trade-off as a sequential "game" between two players. At each step in the gaming sequence, one player maximizes accuracy and the other player imposes a particular amount of fairness. Fairness, which can be defined in several different ways, translates into a set of linear constraints imposed on accuracy that can also be represented as costs. These fairness-specific costs are weights easily ported to a wide variety of classifiers, including some off-the-shelf software. The technical advances from this work are important, but as before, only some kinds of fairness are addressed.

M. Kearns and colleagues (2018) build on the idea of a reduction. They formulate a sequential zero-sum game between a "learner" seeking accuracy and an "auditor" seeking fairness. The algorithm requires users to specify a framework in which groups at risk to unfairness are defined. For example, one might consider all intersections of a set of attributes such as gender, race, and gang membership (e.g., black, male, gang members). The groups that can result are less coarse than groups defined by a single attribute such as race. Equal fairness is imposed over *all* such groups. Because the number of groups can be very large, there would ordinarily be difficult computational problems. However, the reduction leads to a practical algorithm that can be seen as a form of weighting.

Fairness can be defined at the level of individuals (Dwork et al. 2012; Joseph et al. 2016). The basic idea is that similarly situated individuals should be treated similarly. Berk and colleagues (2017) propose a logistic regression classifier with a conventional complexity regularizer and a fairness regularizer operating at the individual level. One of their fairness regularizers evaluates the difference between fitted probabilities for individuals across protected classes. For example, the fitted probabilities of an arrest for black offenders are compared offender by offender to the fitted probabilities of an arrest for white offenders. Greater disparities imply less fairness. Also considered is offender by offender actual outcomes (e.g., arrest or not). Disparities in the fitted probabilities are given more weight if the actual outcome is the same. Ridgeway and Berk (2017) apply a similar individual approach to stochastic gradient boosting. However, mapping individual definitions of unfairness to group-based definitions has yet to be effectively addressed.

## Postprocessing

Postprocessing means that after $h(L, S)$ is applied, its performance is adjusted to make it more fair. To date, perhaps the best example of this approach draws on the idea of random reassignment of the class label previously assigned by $h(L, S)$ (Hardt et al. 2016). Fairness, called "equalized odds," requires that the fitted outcome classes (e.g., high risk or low risk) are independent of protected class membership, conditioning on the actual outcome classes. The requisite information is obtained from the rows of a confusion table and, therefore, represents classification accuracy, not prediction accuracy. There is a more restrictive definition called "equal opportunity" requiring such fairness only for the more desirable of the two outcome classes.[26]

For a binary response, some cases are assigned a value of 0 and some assigned a value of 1. To each is attached a probability of switching from a 0 to a 1 or from a 1 to a 0 depending in whether a 0 or a 1 is the outcome assigned by $\hat{f}^{*}(L, S)$. These probabilities can differ from one another and both can differ across different protected group categories. Then, there is a linear programming approach to minimize the classification errors subject to one of the two fairness constraints. This is accomplished by the values chosen for the various probabilities of reassignment. The result is a $\hat{f}^{*}(L, S)$ that achieves conditional procedure accuracy equality.

The implications of this approach for other kinds of fairness are not clear, and conditional use accuracy (i.e., equally accurate predictions) can be a casualty. It is also not clear how best to build in the relative costs of false negatives and false positives. And, there is no doubt that accuracy will suffer more when the probabilities of reassignment are larger. Generally, one would expect to have overall classification accuracy comparable to that achieved for the protected group category for which accuracy is the worst. Moreover, the values chosen for the reassignment probabilities will need to be larger when the base rates across the protected group categories are more disparate. In other words, when conditional procedure accuracy equality is most likely to be in serious jeopardy, the damage to conditional procedure accuracy will be the greatest. More classification errors will be made; more 1s will be treated as 0s and more 0s will be treated as 1s. A consolation may be that everyone will be *equally* worse off.

## Making Fairness Operational

It has long been recognized that efforts to make criminal justice decisions more fair must resolve a crucial auxiliary question: equality with respect to

what benchmark (Blumstein et al. 1983)? To take an example from today's headlines (Corbett-Davies et al. 2017; Salman, Coz, and Johnson 2016), should the longer prison terms of black offenders be on the average the same as the shorter prison terms given to white offenders or should the shorter prison terms of white offenders be on the average the same as the longer prison terms given to black offenders? Perhaps one should split the difference? Fairness by itself is silent on the choice, which would depend on views about the costs and benefits of incarceration in general. All of the proposed corrections for unfairness we have found are agnostic about what the target outcome for fairness should be. If there is a policy preference, it should be built into the algorithm, perhaps as additional constraints or through an altered loss function. For instance, if mass incarceration is the dominant concern, the shorter prison terms of white offenders might be a reasonable fairness goal for both whites and blacks.[27]

We have been emphasizing binary outcomes, and the issues are much the same. For example, whose conditional use accuracy should be the policy target? Should the conditional use accuracy for male offenders or female offenders become the conditional use accuracy for all? An apparent solution is to choose as the policy target the higher accuracy. But that ignores the consequences for the false negative and false positive rates. By those measures, an undesirable benchmark might result. The benchmark determination has made trade-offs more complicated, and some kind of policy balance would need to be found.

## *Future Work*

Corrections for unfairness combine technical challenges with policy challenges. We have currently no definitive responses to either. Progress will likely come in many small steps beginning with solutions from tractable, highly stylized formulations. One must avoid vague or unjustified claims or rushing these early results into the policy arena. Because there is a large market for solutions, the temptations will be substantial. At the same time, the benchmark is current practice. By that standard, even small steps, imperfect as they may be, can in principle lead to meaningful improvements in criminal justice decisions. They just need to be accurately characterized.

But even these small steps can create downstream difficulties. The training data used for criminal justice algorithms necessarily reflect past practices. Insofar as the algorithms affect criminal justice decisions, existing training data may be compromised. Current decisions are being made differently. It will be important, therefore, for existing algorithmic results to be

regularly updated using the most recent training data. Some challenging technical questions follow. For example, is there a role for online learning? How much historical data should be discarded as the training data are revised? Should more recent training data be given more weight in the analysis? But one can imagine a world in which algorithms improve criminal justice decisions, and those improved criminal justice decisions provide training data for updating the algorithms. Over several iterations, the accumulated improvements might be dramatic.

## A Brief Empirical Example of Fairness Trade-offs with In-processing

There are such stark differences between men and women with respect to crime that cross-gender comparisons allow for relatively simple and instructive discussions of fairness. However, they also convey misleading impressions of the role of fairness in general. The real world can be more complicated and subtle. To illustrate, we draw on some ongoing work being undertaken for a jurisdiction concerned about racial bias that could result from release decisions at arraignment. The brief discussion to follow will focus on in-processing adjustments for bias. Similar problems can arise for preprocessing and postprocessing.

At a preliminary arraignment, a magistrate must decide whom to release awaiting that offender's next court appearance. One factor considered, required by statute, is an offender's threat to public safety. A forecasting algorithm currently is being developed, using the machine learning procedure random forests, to help in the assessment of risk. We extract a simplified illustration from that work for didactic purposes.

The training data are comprised of black and white offenders who had been arrested and arraigned. As a form of in-processing, random forests was applied separately to black and white offenders. Accuracy was first optimized for whites. Then, the random forests application to the data for blacks were tuned so that conditional use accuracy was virtually same as for whites. The tuning was undertaken using stratified sampling as each tree in the forest was grown, the outcome classes as strata. This is effectively the same as changing the prior distribution of the response and alters each tree. All of the output can change as a result. This is very different from trying to introduce more fairness in the algorithmic output alone.

Among the many useful predictors were age, prior record, gender, date of the next most recent arrest, and the age at which an offender was first charged as an adult. Race and residence zip code were not included as predictors.[28]

**Table 16.** Fairness Analysis for Black and White Offenders at Arraignment Using as an Outcome an Absence of Any Subsequent Arrest for a Crime of Violence (13,396 Blacks; 6,604 Whites).

| Race | Base Rate | Conditional Use Accuracy | False Negative Rate | False Positive Rate |
|---|---|---|---|---|
| Black | .89 | .93 | .49 | .24 |
| White | .94 | .94 | .93 | .02 |

Two outcome classes are used for this illustration: within 21 months of arraignment, an arrest for a crime of violence (i.e., a failure) or no arrest for a crime of violence (i.e., a success). We use these two categories because should a crime of violence be predicted at arraignment, an offender would likely be detained. For other kinds of predicted arrests, an offender might well be freed or diverted into a treatment program. A prediction of no arrest might well lead to a release.[29] A 21-month follow up may seem inordinately lengthy, but in this jurisdiction, it can take that long for a case to be resolved.[30]

Table 16 provides the output that can be used to consider the kinds of fairness commonly addressed in the existing criminal justice literature. Success base rates are reported on the far left of the table, separately for blacks and whites: .89 and .94 respectively. For both, the vast majority of offenders are not arrested for a violent crime, but blacks are more likely to be arrested for a crime of violence after a release. It follows that the white rearrest rate is .06, and the black rearrest rate is .11, nearly a 2 to 1 difference.

For this application, we focus on the probability that when the absence of an arrest for a violent crime is forecasted, the forecast is correct. The two different applications of random forests were tuned so that the probabilities are virtually the same: .93 and .94. There is conditional use accuracy equality, which some assert is a necessary feature of fairness.

But as already emphasized, except in very unusual circumstances, there are trade-offs. Here, the false negative and false positive rates vary dramatically by race. The false negative rate is much higher for whites so that violent white offenders are more likely than violent black offenders to be incorrectly classified as nonviolent. The false positive rate is much higher for blacks so that nonviolent black offenders are more likely than nonviolent white offenders to be incorrectly classified as violent. Both error rates mistakenly inflate the relative representation of blacks predicted to be violent. Such differences can support claims of racial injustice. In this application, the trade-off between two different kinds of fairness has real bite.

One can get another perspective on the source of the different error rates from the ratios of false negatives and false positives. From the cross-tabulation (i.e., confusion table) for blacks, the ratio of the number of false positives to the number of false negatives is a little more than 4.2. One false negative is traded for 4.2 false positives. From the cross-tabulation for whites, the ratio of the number of false *negatives* to the number of false *positives* is a little more than 3.1. One false positive is traded for 3.1 false negatives. For blacks, false negatives are especially costly so that the algorithm works to avoid them. For whites, false positives are especially costly so that the algorithm works to avoid them. In this instance, the random forest algorithm generates substantial treatment inequality during in-processing while achieving conditional use accuracy equality.

With the modest difference in base rates, the large difference in treatment equality may seem strange. But recall that to arrive at conditional use accuracy equality, random forests were grown and tuned separately for blacks and whites. For these data, the importance of specific predictors often varied by race. For example, the age at which offenders received their first charge as an adult was a very important predictor for blacks but not for whites. In other words, the *structure* of the results was rather different by race. In effect, there was one $h_B(L, S)$ for blacks and another $h_W(L, S)$ for whites, which can help explain the large racial differences in the false negative and false positive rates. With one exception (Joseph et al. 2016), different fitting structures for different protected group categories have to our knowledge not been considered in the technical literature, and it introduces significant fairness complications (Zliobaite and Custers 2016).[31]

In summary, Table 16 illustrates well the formal results discussed earlier. There are different kinds of fairness that in practice are incompatible. There is no technical solution without some price being paid. How the trade-offs should be made is a political decision.

## Conclusions

In contrast to much of the rhetoric surrounding criminal justice risk assessments, the problems can be subtle, and there are no easy answers. Except in stylized examples, there will be trade-offs. These are mathematical facts subject to formal proofs (Chouldechova 2017; Kleinberg et al. 2016). Denying that these trade-offs exist is not a solution. And in practice, the issues can be even more complicated, as we have just shown.

Perhaps the most challenging problem in practice for criminal justice risk assessments is that different base rates are endemic across protected group

categories. There is, for example, no denying that young men are responsible for the vast majority of violent crimes. Such a difference can cascade through fairness assessments and lead to difficult trade-offs.

Criminal justice decision makers have begun wrestling with the issues. One has to look no further than the recent ruling by the Wisconsin Supreme Court, which upheld the use of one controversial risk assessment tool (i.e., COMPAS) as one of many factors that can be used in sentencing (State of Wisconsin v. Eric L. Loomis, Case # 2915AP157-CR). Fairness matters. So does accuracy.

There are several potential paths forward. First, criminal justice risk assessments have been undertaken in the United States since the 1920s (Borden 1928; Burgess 1928). Recent applications of advanced statistical procedures are just a continuation of long-term trends that can improve transparency and accuracy, especially compared to decisions made solely by judgment (Berk and Hyatt 2015). They also can improve fairness. But categorical endorsements or condemnations serve no one.

Second, as statistical procedures become more powerful, especially when combined with "big data," the various trade-offs need to be explicitly represented and available as tuning parameters that can be easily adjusted. Such work is underway, but the technical challenges are substantial. There are conceptual challenges as well, such as arriving at measures of fairness with which trade-offs can be made. There too, progress is being made.

Third, in the end, it will fall to stakeholders—not criminologists, not statisticians, and not computer scientists—to determine the trade-offs. How many unanticipated crimes are worth some specified improvement in conditional use accuracy equality? How large an increase in the false negative rate is worth some specified improvement in conditional use accuracy equality? These are matters of values and law, and ultimately, the political process. They are not matters of science.

Fourth, whatever the solutions and compromises, they will not come quickly. In the interim, one must be prepared to seriously consider modest improvements in accuracy, transparency, and fairness. One must not forget that current practice is the operational benchmark (Salman et al. 2016). The task is to try to improve that practice.

Finally, one cannot expect any risk assessment tool to reverse centuries of racial injustice or gender inequality. That bar is far too high. But, one can hope to do better.

## Authors' Note

Caroline Gonzalez Ciccone provided very helpful editing suggestions.

## Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

## Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## Notes

1. Many of the issues apply to actuarial methods in general about which concerns have been raised for some time (Messinger and Berk 1987; Feeley and Simon 1994).
2. An algorithm is not a model. An algorithm is a sequential set of instructions for performing some task. When a checkbook is balanced, an algorithm is being applied. A model is an algebraic statement about how the world works. In statistics, often it represents how the data were generated.
3. Similar reasoning is often used in the biomedical sciences. For example, a success can be a diagnostic test that identifies a lung tumor.
4. Language here can get a little murky because the most accurate term depends on the use to which the algorithmic output will be put. We use the term "predicted" to indicate when one is just referring to fitted values (i.e., in training data) and also the when one is using fitted values to make a forecast about an outcome that has not yet occurred.
5. We proceed in this manner because there will be clear links to fairness. There are many other measures from such a table for which this is far less true. Powers (2011) provides an excellent review.
6. There seems to be less naming consistency for these of kinds errors compared to false negatives and false positives. Discussions in statistics about generalization error (Hastie et al. 2009:section 7.2) can provide one set of terms whereas concerns about errors from statistical tests can provide another. In neither case, moreover, is the application to confusion tables necessarily natural. Terms like the "false discovery rate" and the "false omission rate," or "Type II" and "Type I" errors can be instructive for interpreting statistical tests but build in content that is not relevant for prediction errors. There is no null hypothesis being tested.
7. For many kinds of criminal justice decisions, statutes require that decision makers take "future dangerousness" into account. The state of Pennsylvania, for example, requires sentencing judges to consider future dangerousness. Typically, the means by which such forecasts are made is unspecified and in practice, can

depend on the experience, judgment, and values of the decision maker. This might be an example of a sensible calibration benchmark.

8. The binary response might be whether an inmate is reported for serious misconduct such as an assault on a guard or another inmate.
9. How a class of people becomes protected can be a messy legislative and judicial process (Rich 2014). Equally messy can be how to determine when an individual is a member of a particular protected class. For this article, we take as given the existence of protected groups and clear group membership.
10. The IID requirement can be relaxed somewhat (Rosenblatt 1956; Wu 2005). Certain kinds of dependence can be tolerated. For example, suppose the dependence between any pair of observations declines with the distance between the two observations and at some distance of sufficient size becomes independence. A central limit theorem then applies. Perhaps the most common examples are found when data are arrayed in time. Observations that are proximate to one another may be correlated, but with sufficient elapsed time become uncorrelated. These ideas can apply to our discussion and permit a much wider range of credible applications. However, the details are beyond the scope of this article.
11. A joint probability distribution is essentially an abstraction of a high-dimensional histogram from a finite population. It is just that the number of observations is now limitless, and there is no binning. As a formal matter, when all of the variables are continuous, the proper term is a joint density because densities rather than probabilities are represented. When the variables are all discrete, the proper term is a joint probability distribution because probabilities are represented. When one does not want to commit to either or when some variables are continuous and some are discrete, one commonly uses the term joint probability distribution. That is how we proceed here.
12. Science fiction aside, one cannot assume that even the most powerful machine learning algorithm currently available with access to all of the requisite predictors will "learn" the true response surface. And even if it did, how would one know? To properly be convinced, one would already have to know the true response surface, and then, there would be no reason to estimate it (Berk 2016a:section 1.4).
13. The normal equations, which are the source of the least squares solution in linear regression, are a special case.
14. We retain $S$ in the best approximation even though it represents protected groups. The wisdom of proceeding in this manner is considered later when fairness is discussed. But at the very least, no unfairness can be documented unless $S$ is included in the data.
15. There can be challenges in practice if, for example, $h(L, S)$ is tuned with training data. Berk and his colleagues (2018) provide an accessible discussion.

16. The meaning of "decision" can vary. For some, it is assigning an outcome class to a numeric risk score. For others, it is a concrete, behavioral action taken with the information provided by a risk assessment.
17. Accuracy is simply (1 − error), where error is a proportion misclassified or the proportion forecasted incorrectly.
18. Dieterich and his colleagues (2016:7) argue that overall there is accuracy equity because "the AUCs obtained for the risk scales were the same, and thus equitable, for blacks and whites." The AUC depends on the true positive rate and false positive rate, which condition on the known outcomes. Consequently, it differs formally from overall accuracy equality. Moreover, there are alterations of the AUC that can lead to more desirable performance measures (Powers 2011).
19. One of the two outcome classes is deemed more desirable, and that is the outcome class for which there is conditional procedure accuracy equality. In criminal justice settings, it can be unclear which outcome class is more desirable. Is an arrest for burglary more or less desirable than an arrest for a straw purchase of a firearm? But if one outcome class is recidivism and the other outcome class is no recidivism, equality of opportunity refers to conditional procedure accuracy equality for those who did not recidivate.
20. Chouldechova builds on numeric risk scores. A risk instrument is said to be well calibrated when predicted probability of the preferred outcome (e.g., no arrest) is the same for different protected group classes at each risk score value—or binned versions of those values. Under these circumstances, it is possible for a risk instrument to have predictive parity but not be well calibrated. For reasons that are for this article peripheral, both conditions are the same for a confusion table with a binary outcome. For Kleinberg et al. (2016:4), a risk instrument that is well calibrated requires a bit more. The risk score should perform like a probability. It is not apparent how this would apply to a confusion table.
21. One can turn the problem around and consider the degree to which individuals who have the same binary outcome (e.g., an arrest) have similar predicted outcomes and whether the degree of similarity in predicted outcomes varies by protected class membership (Berk, Heidari et al. 2017; Ridgeway and Berk 2017).
22. For example, machine learning algorithms usually are inductive. They engage in automated "data snooping," and an empirical determination of tuning parameter values exacerbates the nature and extent of the overfitting. Consequently, one should not apply the algorithm anew to the test data. Rather, the algorithmic output from the training data is taken as given, and fitted values using the test data are obtained.
23. When base rates are the same in this example, one perhaps could achieve perfect fairness while also getting perfect accuracy. The example doesn't have enough

information to conclude that the populations aren't separable. But that is not the point we are trying to make.

24. The numbers in each cell assume for arithmetic simplicity that the counts come out exactly as they would in a limitless number of realizations. In practice, an assignment probability of .30 does not require exact cell counts of 30 percent.
25. Although statistical parity has not figured in these illustrations, changing the base rate negates it.
26. In criminal justice applications, determining which outcome is more desirable will often depend on which stakeholders you ask.
27. Zliobaite and Custers (2016) raise related concerns for risk tools derived from conventional linear regression for lending decisions.
28. Because of racial residential patterns, zip code can be a strong proxy for race. In this jurisdiction, stakeholders decided that race and zip code should not be included as predictors. Moreover, because of separate analyses for whites and blacks, race is a constant within each analysis.
29. Actually, the decision is more complicated because a magistrate must also anticipate whether an offender will report to court when required to do so. There are machine learning forecasts being developed for failures to appear, but a discussion of that work is well beyond the scope of this article.
30. The project is actually using four outcome classes, but a discussion of those results complicates things unnecessarily. They require a paper of their own.
31. There are a number of curious applications of statistical procedures in the Zliobaite and Custers paper (e.g., propensity score matching treating gender like an experimental intervention despite it being a fixed attribute). But the concerns about fairness when protected groups are fitted separately are worth a serious read.

## References

Agarwal, A., A. Beygelzimer, M. Dudk, J. Langford, and H. Wallach. 2018. "A Reductions Approach to Fair Classification." Preprint available at https://arxiv.org/abs/1803.02453.

American Diabetes Association. 2018. "Statistics about Diabetes" (http://www.diabetes.org/diabetes-basics/statistics/).

Angwin, J, J. Larson, S. Mattu, and L. Kirchner. 2016. "Machine Bias." (https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing).

Barocas, S., and A. D. Selbst. 2016. "Big Data's Disparate Impact." *California Law Review* 104:671-732.

Berk, R. A. 2008. "The Role of Race in Forecasts of Violent Crime." *Race and Social Problems* 1:231-242.

Berk, R. A. 2012. *Criminal Justice Forecasts of Risk: A Machine Learning Approach*. New York: Springer.

Berk, R. A. 2016a. *Statistical Learning from a Regression Perspective*. 2nd ed. New York: Springer.

Berk, R. A. 2016b. "A Primer on Fairness in Criminal Justice Risk Assessments." *The Criminologist* 41(6):6-9.

Berk, R. A. 2017. "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology* 13(2):193-216.

Berk, R. A. and J. Bleich. 2013. "Statistical Procedures for Forecasting Criminal Behavior: A Comparative Assessment." *Journal of Criminology and Public Policy* 12(3):513-544.

Berk, R. A., L. Brown, A. Buja, E. George, and L. Zhao. 2017. "Working with Misspecified Regression Models." *Journal of Quantitative Criminology*.

Berk, R. A., L. Brown, E. George, A. Kuchibhotla, W. Weijie Su, and L. Zhao. 2018. "Assumption Lean Regression." Working Paper, Department of Statistics, University of Pennsylvania.

Berk, R. A. and J. de Leeuw. 1999. "An Evaluation of California's Inmate Classification System Using a Generalized Regression Discontinuity Design." *Journal of the American Statistical Association* 94(448):1045-1052.

Berk, R. A., H. Heidari, S. Jabbari, M. Joseph, M. Kearns, J. Morganstern, S. Neel, and A. Roth. 2017. "A Convex Framework for Fair Regression." *arXiv" 1706. 02409v1 [cs.LG]*.

Berk, R. A. and J. Hyatt. 2015. "Machine Learning Forecasts of Risk to Inform Sentencing Decisions." *The Federal Sentencing Reporter* 27(4):222-228.

Berk, R. A. and S. B. Sorenson. 2016. "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions." *Journal of Empirical Legal Studies* 31(1):94-115.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. New York: Springer.

Blumstein, A., J. Cohen, S. E. Martin, and M. H. Tonrey. 1983. *Research on Sentencing: The Search for Reform, Volume I*. Washington, DC: National Academy Press.

Borden, H. G. 1928. "Factors Predicting Parole Success." *Journal of the American Institute of Criminal Law and Criminology* 19: 328-336.

Brennan, T. and W. L. Oliver. 2013. "The Emergence of Machine Learning Techniques in Criminology." *Criminology and Public Policy* 12(3):551-562.

Buja, A., R. A. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang. 2018a. "Models as Approximations, Part I—A Conspiracy of Nonlinearity and Random Regressors in Linear Regression." *Statistical Science*, forthcoming with discussion.

Buja, A., R. A. Berk, L. Brown, E. George, E. Pitkin, M. Traskin, L. Zhao, and K. Zhang. 2018b. "Models as Approximations, Part II—A Conspiracy of Nonlinearity and Random Regressors in Linear Regression." *Statistical Science*, forthcoming with discussion.

Burgess, E. M. 1928. "Factors Determining Success or Failure on Parole." Pp. 205-249 in *The Working of the Indeterminate Sentence Law and the Parole System in Illinois*, edited by A. A. Bruce, A. J. Harno, E. W Burgess, and E. W. Landesco. Springfield, IL: State Board of Parole.

Calmon, F. P., D. Wei, K. N. Ramamurthy, and K. R. Varshney. 2017. "Optimizing Data Pre-Processing for Discrimination Prevention." arXiv: 1704.03354v1 [stat. ML].

Cohen, A. 2012. "Wrongful Convictions: A New Exoneration Registry Tests Stubborn Judges." *The Atlantic*, May, 21.

Corbett-Davies, S., E. Pierson, A. Feller, S. Goel, and A. Hug. 2017. "Algorithmic Decision Making and Cost of Fairness." *KDD '17: Proceedings of the 23 rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Chouldechova, A. 2017. "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. arXiv:1703.00056v1 [stat.AP].

Crawford, K. 2016. "Artificial Intelligence's White Guy Problem." *New York Times*, Sunday Review, June 25.

Demuth, S. 2003. "Racial and Ethnic Differences in Pretrial Release Decisions and Outcomes: A Comparison of Hispanic, Black and White Felony Arrestees." *Criminology* 41:873-908.

Dieterich, W., C. Mendoza, and T. Brennan. 2016. *"COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity."* Northpoint.

Doleac, J. and M. Stevenson. 2016. "Are Criminal Justice Risk Assessment Scores Racist?" Brookings Institute. (https://www.brookings.edu/blog/up-front/2016/08/22/are-criminal-risk-assessment-scores-racist/).

Dwork, C., Y. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. "Fairness through Awareness." In *Proceedings of the 3rd Innovations of Theoretical Computer Science*, 214-226.

Feeley, M. and J. Simon. 1994. "Actuarial Justice: The Emerging New Criminal Law." Pp. 173-201 in *The Futures of Criminology*, edited by D. Nelken. London, UK: Sage.

Feldman, M., S. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubrtamanian. 2015. "Certifying and Removing Disparate Impact." In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259-268.

Ferguson, A. G. 2015. "Big Data and Predictive Reasonable Suspicion." *University of Pennsylvania Law Review* 163(2):339-410.

Freedman, D. A. 1981. "Bootstrapping Regression Models." *Annals of Statistics* 9(6): 1218-1228.

Friedler, S. A., C. Scheidegger, and S. Venkatasubramanian. 2016. "On The (Im)possibility of Fairness)." axXiv1609.07236v1 [cs.CY].

Hajian, S. and J. Domingo-Ferrer. 2013. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining." *IEEE transactions on knowledge and data engineering* 25(7):1445-1459.

Harcourt, B. W. 2007. "Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age." Chicago, IL: University of Chicago Press.

Hardt, M., E. Price, and N. Srebro. 2016 "Equality of Opportunity in Supervised Learning." Pp. 3315-23 in *Equality of Opportunity in Supervised Learning* Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5–10, 2016, edited by D. D. Lee, Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. (eds.). Barcelona, Spain.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

Hamilton, M. 2016. "Risk-needs Assessment: Constitutional and Ethical Challenges." *American Criminal Law Review* 52(2):231-292.

Hyatt, J. M., L. Chanenson, and M. H. Bergstrom. 2011. "Reform in Motion: The Promise and Profiles of Incorporating Risk Assessments and Cost-Benefit Analysis into Pennsylvania Sentencing." *Duquesne Law Review* 49(4):707-749.

Janssen, M. and G. Kuk. 2016. "The Challenges and Limits of Big Data Algorithms in Technocratic Governance." *Government Information Quarterly* 33:371-377.

Johndrow, J.E. and K. Lum. 2017. "An Algorithm for Removing Sensitive Information: Application to Race-independent Recidivism Prediction." arXIV:1703. 049557v1 [stat.AP].

Johnson, K. D., D. P. Foster, and R. A. Stine. 2016. "Impartial Predictive Modeling: Ensuring Fairness in Arbitrary Models." arXIV:1606.00528v1 [stat.ME].

Joseph, M., M. Kearns, J. H. Morgenstern, and A. Roth. 2016. Pp. 325-33 in *Fairness in Learning: Classic and Contextual Bandits*. Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. (eds.). Barcelona, Spain.

Kamiran, F. and T. Calders. 2009. "Classifying Without Discrimination." *2009 2nd International Conference on Computer, Control and Communication*, IC4 2009.

Kamiran, F. and T. Calders. 2012. "Data Preprocessing Techniques for Classification without Discrimination." *Knowledge Information Systems* 33:1-33.

Kamishima, T., S. Akaho, and J. Sakuma. 2011. "Fairness-aware Learning through a Regularization Approach." *Proceedings of the 3rd IEEE International Workshop on Privacy Aspects of Data Mining*.

Kearns, M. J. 1994. An Introduction to Computational Learning Theory. Cambridge, UK: The MIT Press.

Kearns, M., S. Neel, A. Roth, and S. Wu. 2018. "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." Preprint https://arxiv.org/abs/1711.05144.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. 2017. "Human Decisions and Machine Predictions." NBER Working paper 23180. National Bureau of Economic Resaerch.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2016. "Inherent Trade-Offs in Fair Determination of Risk Scores." arXiv: 1609.05807v1 [cs.LG].

Kroll, J. A., J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. 2017. "Accountable Algorithms." *University of Pennsylvania Law Review* 165(3):633-705.

Liu, Y. Y., M. Yang, M. Ramsay, X. S. Li, and J. W. Cold. 2011. "A Comparison of Logistic Regression, Classification and Regression Trees, and Neutral Networks Model in Predicting Violent Re-Offending." *Journal of Quantitative Criminology* 27:547-573.

Messinger, S. L. and R. A. Berk. 1987. "Dangerous People: A Review of the NAS Report on Career Criminals." *Criminology* 25(3):767-781.

National Science and Technology Council. 2016. "Preparing for the Future of Artificial Intelligence." Executive of the President, National Science and Technology Council, Committee on Technology.

Pedreschi, D., S. Ruggieri, and F. Turini. 2008. "Discrimination-aware Data Mining." KDD2008, August 24–27, 2008, Las Vegas, Nevada, USA.

Pew Center of the States, Public Safety Performance Project. 2011. "Risk/Needs Assessment 101: Science Reveals New Tools to Manage Offenders." The Pew Center of the States (http://www.pewcenteronthestates.org/publicsafety).

Pleiss, G., M. Raghavan, F. Wi, J. Kleinberg, and K.Q. Weinberger. 2017. "On Fairness and Calibration." arXiv:1709.02012v1 [cs.LG].

Powers, D. M. W. 2011. "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation." *Journal of Machine Learning Technologies* 2(1):37-63.

Rich, C. G. 2014. "Elective Race: Recognizing Race Discrimination in the Era of Racial Self-identification." *Georgetown Law Journal* 102:1501-1572.

Ridgeway, G. 2013a. "The Pitfalls of Prediction." *NIJ Journal* (271):34-40.

Ridgeway, G. 2013b "Linking Prediction to Prevention." *Criminology and Public Policy* 12(3):545-550.

Ridgeway, G. and R. Berk. 2017. "Fair Gradient Boosting." Working paper, Department of Criminology, University of Pennsylvania.

Rhodes, W. 2013. "Machine Learning Approaches as a Tool for Effective Offender Risk Prediction." *Criminology and Public Policy* 12(3):507-510.

Rosenblatt, M. 1956. "A Central Limit Theorem and A Strong Mixing Condition" *Proceeding of the National Academy of Sciences* 42:43-47.

Salman, J., E. L. Coz, and E. Johnson. 2016. "Florida's Broken Sentencing System. Sarasota Herald Tribune (http://projects.heraldtribune.com/bias/sentencing/).

Starr, S. B. 2014a. "Sentencing by the Numbers." *New York Times op-ed*, August 10, 2014.

Starr, S. B. 2014b. "Evidence-based Sentencing and The Scientific Rationalization of Discrimination." *Stanford Law Review* 66:803-872.

Tonry, M. 2014. "Legal and Ethical Issues in The Prediction of Recidivism." *Federal Sentencing Reporter* 26(3):167-176.

White, H. 1980. "Using Least Squares to Approximate Unknown Regression Functions." *International Economic Review* 21(1):149-170.

Wu, W.B. 2005. "Nonlinear System Theory: Another Look at Dependence." *PNAS* 102(40):14150-14154.

Zliobaite, I. and B. Custers. 2016. "Using Sensitive Personal Data May Be Necessary for Avoiding Discrimination in Data-Driven Decision Models." *Artificial Intelligence and the Law* 24(2):183-201.

## Author Biographies

**Richard Berk** is a professor in the Department of Criminology and the Department of Statistics at the University of Pennsylvania.

**Hoda Heidari** is a postdoctoral fellow at ETH Zurich.

**Shahin Jabbari** is a graduate student in the Department of Computer and Information Science at the University of Pennsylvania.

**Michael Kearns** is a professor in the Department of Computer and Information Science at the University of Pennsylvania.

**Aaron Roth** is a professor in the Department of Computer and Information Science at the University of Pennsylvania.