ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Fair Risk Algorithms

Richard A. Berk,[1,2] Arun Kumar Kuchibhotla,[3] and Eric Tchetgen Tchetgen[2]

[1]Department of Criminology, University of Pennsylvania, Philadelphia, Pennsylvania, USA; email: berkr@sas.upenn.edu.edu

[2]Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Annu. Rev. Stat. Appl. 2023. 10:165–87

First published as a Review in Advance on October 7, 2022

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-033021-120649

Keywords

fairness, discrimination, algorithms, risk assessment, criminal justice, machine learning, optimal transport, conformation prediction

Abstract

Machine learning algorithms are becoming ubiquitous in modern life. When used to help inform human decision making, they have been criticized by some for insufficient accuracy, an absence of transparency, and unfairness. Many of these concerns can be legitimate, although they are less convincing when compared with the uneven quality of human decisions. There is now a large literature in statistics and computer science offering a range of proposed improvements. In this article, we focus on machine learning algorithms used to forecast risk, such as those employed by judges to anticipate a convicted offender's future dangerousness and by physicians to help formulate a medical prognosis or ration scarce medical care. We review a variety of conceptual, technical, and practical features common to risk algorithms and offer suggestions for how their development and use might be meaningfully advanced. Fairness concerns are emphasized.

1. INTRODUCTION

In this article, we consider algorithms used to help inform human decisions that are contingent on future events. Settings for these decisions can include mortgage lending, parole releases from prison, premium determination for flood insurance, employee hiring, medical prognoses, enforcement of environmental regulations, local tornado warnings, and many more (Coglianese & Lai 2022). In each case, the decision is informed by algorithmic forecasts of possible occurrences in the future, in the much same spirit as actuarial determinations of risk. Ideally, such forecasts are sufficiently accurate and made through acceptably transparent procedures. Fairness matters, too, and can dictate the manner in which a procedure is received, which in turn can affect whether that procedure is actually deployed (Starr 2014).

Unless algorithmic forecasts are perceived by stakeholders as fair, accuracy and transparency may not matter much. Usual practice specifies an objective function that the algorithm is designed to optimize, commonly related to prediction accuracy. Cross-entropy is one example. But minimizing the loss usually has collateral consequences, with unfairness perhaps the most important and controversial. Ideally, the algorithm should be sufficiently accurate but also sufficiently fair.

We focus on fairness. Page constraints preclude much discussion of accuracy and transparency. We also do not address backward-looking algorithms used in fraud detection, facial recognition, and other retrospective tasks. Was a particular sales transaction, for instance, made by the owner of the credit card? Was the person identified by an algorithm leaving a crime scene the same person arrested for that crime (Singer & Metz 2019)? Fairness certainly can be an issue, but accurate classification, not forecasting, is the goal.

Before turning to risk algorithms themselves, we set the stage by drawing on a famous study by Bickel et al. (1975) examining possible gender bias in graduate admissions at the University of California, Berkeley. In the 1970s, claims were made that at Berkeley, a smaller proportion of women than men were admitted for graduate study. Because academic merit was supposed to be the dominant admissions metric, fairness concerns were raised. On closer inspection, the researchers learned that applicants to more competitive graduate programs were more likely to be women than men, and that the proportions of men and women admitted to these programs were actually comparable. But because the admissions rates were lower for the more competitive programs, women applicants had a less chance of being admitted to the Berkeley graduate programs overall.

Two conclusions followed. Over the full graduate program, men and women on the average were not similarly situated. The particular programs to which they applied often differed in their competitiveness. In addition, the competitiveness of an applicant's program was seen as a legitimate factor affecting graduate admissions, meaning that any claims of unfair treatment overall were moot. The empirical admission disparity was not discriminatory.

Bickel et al.'s (1975) analysis centered on existing admissions outcomes to evaluate potential difficulties with Berkeley's graduate admissions procedures. But at least implicitly, there was the prospect of altering admissions procedures in the future if there was strong evidence of gender unfairness. Consequently, evaluation of any proposed admissions reforms would require credible statistical and causal inference. Would fairness be achieved if particular admissions reforms were implemented? Could uncertainty be properly represented in that context? Because any reforms would affect Berkeley applicants in the future, available data collected under existing admissions practices might be insufficiently informative.

Consider two extreme, but fair, admissions procedures. One could have equal gender admissions proportions overall by admitting no applicants or by admitting all applicants. Academic merit would be irrelevant, and the mix of admitted students would change dramatically under

either scenario. The first procedure would make everyone who would have been admitted worse off. The second would make everyone who would not have been admitted better off. Under the first procedure, Berkeley would have no incoming graduate students. Under the second procedure, the incoming class would surely be too large. However, credible projections, essential for deciding empirically whether to change admissions procedures, would be extremely difficult to undertake at the time when a reform decision was needed. For example, if all applicants were admitted, would entry-level courses be used to weed out students unable to do the work? Would that be fair? Retention disparities might just reproduce the earlier admissions disparities.

The circumstantial issues would be much the same if a merit-based, fair algorithm rather than a merit-based, fair committee procedure was proposed for making admissions decisions, even if that algorithm were just providing information to a committee. There again would be cascading consequences that would need to be anticipated before an informed decision could be made about whether admissions decisions by algorithm should be adopted. For example, would the algorithm alter the academic capabilities of future admissions cohorts for the better or for the worse? And as before, the only data available on which to base such projections would represent past admissions cohorts selected by prior procedures.

The Berkeley graduate admissions example broadly frames the discussion to follow. But it takes for granted how fairness is defined, how it might be measured, and how it might be achieved. It also uses the term "algorithm" as little more than a procedural placeholder leaving its meaning and operation to the imagination. We address both matters as we examine the prospects for fair risk algorithms.

Fairness is at the center of all that follows. In this setting, accuracy typically means estimates of classification error and forecasting error. Readers wanting a rich consideration of accuracy will find excellent treatments in any number of textbooks, such as those by Bishop (2006), Hastie et al. (2009), Murphy (2012), and Berk (2020). For risk algorithms, transparency commonly addresses the degree to which the features of algorithms that determine computed risk can be easily explained. Thoughtful expositions are difficult to find, but Rudin & Berk (2018) have an applied discussion in criminal justice and healthcare settings, and Coglianese (2021) addresses transparency in the larger context of automated government regulation.

2. ALGORITHMS AND MODELS

Consistent with recent reviews (Mitchell et al. 2021), we concentrate on algorithms used to estimate risk. We pay little attention to models sometimes undertaking similar tasks. The exposition leans heavily on the strengths of algorithms discussed by Breiman (2001) and skirts the weaknesses of models addressed by Buja et al. (2019a,b). In this manner, some ambiguities in the fairness literature can be avoided and our discussion can be held to a reasonable length.

Although the fairness literature in statistics and computer science often uses the terms algorithms and models interchangeably, there are important differences in methods and perspectives. "An algorithm is nothing more than a very precisely specified series of instructions for performing some concrete task" (Kearns & Roth 2020, p. 4). Its usual purpose when assessing risk is to accurately forecast some outcome, typically using supervised learning. For example, a pattern recognition algorithm might be used to forecast future malignancies from lung X-rays. There is no necessary explanatory or causal content in that forecast. Also, the algorithm is not a description of how the data were generated.

In contrast, a model typically is a statement about how the data were generated, intended to facilitate explanation and, sometimes, causal inference. The focus is on the manner in which some empirical phenomenon works, and understanding is commonly the dominant goal.

In part because the goals of algorithms and models can differ substantially, risk algorithms and models are typically governed by different performance criteria. Algorithms are neither right nor wrong. Hence, misspecification is not formally defined. Rather, risk algorithms fit the data in a satisfactory manner or they do not. An algorithm designed to forecast which small businesses will default on a loan is not designed to explain why such enterprises fail. What matters is whether the algorithm provides forecasts that can reduce substantially a lender's financial exposure.

Models are commonly meant to closely approximate some real phenomenon. For instance, understanding the causes of business bankruptcies usually requires structural equation machinery. Consequently, model misspecification is a ubiquitous concern. Models can be right or wrong, and the discourse surrounding models is often dominated by such matters (Watson & Holmes 2016).

The differences between algorithms and models sometimes can be muddy in practice when models are used primarily to classify or forecast, when algorithms are used to make provisional causal claims (Schökopf 2019), or when subject-matter insight and past research are used to help determine the variables an algorithm incorporates (Athey & Imbens 2019). Opaque distinctions between algorithms and models can undermine clarity in concepts and language. For example, a nonparametric regression model can have a foot in both camps.

In summary, whether in practice or in theory, algorithms can be very different from models. Their goals can differ significantly. Performance criteria then can differ significantly as well. Care must be taken not to conflate the two and muddy how one studies fairness.

3. NOTATION AND INITIAL ILLUSTRATION

For better continuity with recent work in the *Annual Review of Statistics and Its Applications*, we employ a notation similar to that used by Mitchell et al. (2021), although we formulate the problem rather differently. We also take some notational liberties for brevity.

Risk algorithms are built and evaluated with historical data. There are N study units, also called cases, in the data indexed by $i = 1, 2, \ldots N$. In discussions of fairness, these units are usually people, but neighborhoods, schools, businesses, or other entities can be appropriate as well. With our focus on algorithms, there is a population of units represented by a joint probability distribution $\mathbb{P}(Y, V)$ containing an outcome Y and unit features V. Sometimes there are several outcomes $Y^{(c)}$, $c = 1, 2, \ldots, C$, although for ease of exposition, our discussion is limited to a single outcome variable. Which random variables are features and which are outcomes has no role in the data generation process; such distinctions are made by the analyst and are often dictated by the problem at hand. The N observations (y_i, v_i), $1 \leq i \leq N$ are realized IID or in an exchangeable manner. Without either of these realization properties, it is difficult to formulate the forecasting enterprise at the center of risk algorithms.

Features V can contain some situation-specific illegitimate features A and some situation-specific legitimate features X. Determining which features are illegitimate and which are legitimate for algorithmic use often will be contentious and depend heavily on the setting. Race commonly is an illegitimate feature when a prison administrator projects in-prison misconduct to determine an inmate's security level (Berk & DeLeeuw 1999), but it is a legitimate feature when a physician uses racial genetic markers to anticipate vulnerability to sepsis (Wu et al. 2021). The same kinds of distinctions apply to modern algorithmic forecasts used to inform prison administrators or physicians.

The outcome Y can be numeric or categorical. By and large, fairness discussions have centered on a categorical Y. Like many other expositions, we treat Y as binary {1, 0} for notational and exposition convenience. For example, $Y = 1$ might denote defaulting on a mortgage loan, and

$Y = 0$ might denote making the required payments. For each unit i, forecasting algorithms seek a suitable approximation of $\mathbb{P}(Y = y_i|V = v_i)$ for $y_i \in \{0, 1\}$. Learning from training data, the approximation ideally is sufficiently accurate, transparent, and fair in test data. Its output can be denoted by $\Psi(V = v_i)$, meant to estimate $\hat{y}_i$ directly or through a threshold on numeric values s_i, typically such as $(s_i = \hat{P}(Y = 1|V = v_i)) > 0.50$. But sometimes, s_i simply is computed as summed features thought to be related to Y, much as one might to construct a Likert Scale. Then a numeric threshold in that metric is applied.

In practice, risk algorithms commonly are built with J units of training data and evaluated with K units of test data: $j = 1, 2, \ldots, J$ and $k = 1, 2, \ldots, K$, with $J + K = N$. Our notation is easily adapted for training data and test data, which is introduced as needed. If the algorithm's performance is acceptable in test data, it may be deployed to forecast the outcome when, for any new unit, there is no outcome label. We denote that unit with $i = N + 1$. The ultimate question in practice is how well the algorithm performs with the $N + 1$ units. The training and test data primarily are a means to that end.

To illustrate the notation and fix some initial ideas, consider briefly a simplified but real example of how parole releases from prison sentences were obtained in one jurisdiction using a risk algorithm (Berk 2017). At the request of a prison parole board responsible for deciding when an inmate's parole petition would be granted, a gradient boosting classifier was trained on data $(y_j^{\text{train}}, x_j^{\text{train}})$ from individuals who in the past had been released on parole from prison. The legitimate covariates x_j^{train} included prior arrests for different kinds of crime, misconduct prison reports, progress in job training, age, and others. One important outcome to be forecasted was future dangerousness, coded as 1 for an arrest for a violent crime within 2 years of release and 0 for no such arrest within 2 years of release.

The training produced for each case $\hat{p}_j(Y^{\text{train}} = 1|V^{\text{train}} = v_j))$. For $\hat{p}_j > 0.50$, the arrest label for a violent crime was assigned. Otherwise, the label for no violent crime arrest was assigned. Evaluation was undertaken using the trained algorithm with test data inserted to yield $\hat{p}_k(Y^{\text{test}} = 1|V^{\text{test}} = v_k)$. Here, too, cases with probabilities larger than 0.50 were assigned the arrest label. Otherwise, a nonarrest label for a violent crime was assigned. The test data results were analyzed for accuracy and fairness using summary measures described shortly.

The algorithm had acceptable accuracy and then was deployed to forecast future dangerousness for each new case that came before the parole board. Each forecasted class, arrested or not arrested, was used to help inform the parole board's discussions. From follow-up data on released inmates, there is some evidence that the accuracy of the parole board's decisions improved, especially for a large mass of equivocal cases toward the middle of the risk distribution. Cases in the tails were easy to accurately forecast by almost any sensible empirical method. There is also some evidence that overall, the crime rate, measured by arrests per capita, declined. One inference was that there were fewer high-risk inmates granted parole. Nevertheless, concerns were raised about possible unfairness as an inadvertent, but troubling, possibility. Currently, use of the algorithm is on hold while the state legislature and stakeholders try to arrive more generally at an agreement about the use of risk algorithms in criminal justice decisions.

Having data on how an algorithm performs when actually put in practice is an important issue later in this article. In this application, however, there were necessarily missing outcome labels for an important subset of inmates. Because the outcome was a postrelease arrest for a violent crime, there was no information on how inmates who were not released would have performed had they been granted parole. Missing outcome labels for important subsets of cases is a widely recognized problem in many settings that can be difficult to solve in real applications (Berk et al. 2016, Kleinberg et al. 2018, Kearns & Roth 2020, Imai et al. 2022).

4. DEFINING FAIRNESS

Definitions of fairness have deep roots in religion, philosophy, and the social sciences (Rawls 2001). The issues are unresolved and, therefore, do not provide clear guidance for algorithm designers. Also unresolved is which stakeholders can participate in how fairness will be defined, in part because of differences in self-interest, world view, and political power.

These complications have led some researchers to dismiss virtually all attempts to impose a priori definitions of fairness on algorithms and propose instead relying on some form of definition-free elicitations of fairness from stakeholders (Jung et al. 2020). This work is in the same spirit as social science research going back decades on public perceptions of crime seriousness and of appropriate punishment (Rossi et al. 1974, Miller et al. 1986). A major concern about such work is a failure to recognize that imposed on the views of stakeholders are binding constraints mandated by precedent, law, and policy that cannot be ignored (Rossi & Berk 1997). Fairness is much more than a beauty contest or, in economics parlance, much more than the fungible, subjective tastes of individuals unfettered by practicalities (Becker 1996).

There is a large literature in statistics, computer science, and law providing many formulations of algorithmic fairness. Some of the ideas are very appealing, but no unifying integration is in the offing. Instructive overviews can be found in works by Berk et al. (2018), Baer et al. (2020), and Mitchell et al. (2021). We do not undertake another broad survey of the existing fairness definitions. We build on some central tendencies and common themes in an effort to advance a revised framework more attuned to the opportunities and obstacles for algorithmic fairness in practice.

Several useful and widely employed kinds of fairness can be defined from a conventional confusion table. For a categorical Y, a confusion table is simply a cross-tabulation of a categorical observed outcome label against the predicted outcome label produced by a classifier. A proper confusion table is constructed from test data used in concert with a classifier, applied earlier to training data, to compute predicted outcome classes.

Importantly, training data and test data do not overlap. They are two independent data sets realized from the same joint probability distribution or a random, disjoint split of an existing data set. In either instance, the test data are uncontaminated by any data-driven specification of an algorithm's structure that would otherwise be a form of data snooping (Kuchibhotla et al. 2021).

Table 1 is a stylized confusion table with letters representing cell counts of cases. Successes (e.g., no arrest on parole) here are treated as positives, and failures (e.g., an arrest on parole) are treated as negatives. The bottom row of the table defines different kinds of fitting error. The far right column defines different kinds of classification error. The former conditions on a fitted outcome, which often serves as a forecasted outcome. The latter conditions on the observed outcome. Each cell and margin value is easily computed with test data.

We let A represent one of the two protected classes a and a^* to which that unit might correspond. In practice, for example, this might represent males and females or Blacks and Whites, respectively. More than two protected classes can be used with no complications beyond more complex notation. The response Y is binary, usually coded as 1 or 0, and might represent,

Table 1 A stylized confusion table from test data for a binary outcome variable

	Fitted failure	Fitted success	Classification error
Observed failure	a (true negative)	b (false positive)	$b/(a+b)$ = false positive rate
Observed success	c (false negative)	d (true positive)	$c/(c+d)$ = false negative rate
Fitted error	$c/(a+c)$	$b/(b+d)$	Overall error = $\frac{(b+c)}{(a+b+c+d)}$

The letters represent counts of the number of observations in the test data. There are two outcomes: success, treated as positive, and failure, treated as negative.

respectively, as before, no arrest for a crime of violence after a parole release and an arrest for a crime of violence after a parole release. Here, too, more than two categories easily can be used, although the analysis becomes more complicated.

With this notation at hand, four kinds of fairness parities directly follow. Note that all of the counts and summary statistics in the table in practice depend on a classifier's fitted values, which are a function of covariates. Building that conditioning into the notation would create unnecessary clutter and perhaps even some confusion. Nevertheless, the role of those covariates is a very important part of our discussion.

- Prediction parity is given by $[\mathbb{P}(\widehat{Y} = y|a) = \mathbb{P}(\widehat{Y} = y|a^*)], y \in \{0, 1\}$, and requires that with $\hat{Y}$ as the fitted outcome, the probability of an arrest after a sentence of probation, for instance, is the same for similarly situated men and women, determined by having the same set of legitimate features $X = x$. Prediction parity is sometimes called demographic parity. When A represents different racial groups, a lack of prediction parity is often identified as a potential contributor to mass incarceration insofar as Black offenders are forecasted by a risk algorithm to have a greater probability of arrest on probation or parole than similarly situated offenders from other racial groups.
- Classification parity is given by $[\mathbb{P}(\hat{Y} = y|Y = y, a) = \mathbb{P}(\hat{Y} = y|Y = y, a^*)], y \in \{0, 1\}$, and requires that the false positive rate is the same for, say, similarly situated men and women, and the false negative rate is the same for similarly situated men and women. Equalized odds is closely related to classification parity but requires that the true positive rate is the same across protected groups and the false positive rate is the same across protected groups (Hardt et al. 2016). Note that for both classification parity and equalized odds, one conditions on the relevant Y class in the test data. This is a source of counterfactual complications considered at some length in Section 8.
- Forecasting accuracy parity is given by $[\mathbb{P}(Y = y|\hat{Y} = y, a) = \mathbb{P}(Y = y|\hat{Y} = y, a^*)], y \in \{0, 1\}$, and requires that fitted accuracy for each outcome class is the same, say, for similarly situated men and women. Because fitted values are often treated as forecasts and real forecasts are central to algorithmic risk assessment, we use the term forecasting accuracy parity. It is sometimes called predictive parity, not to be confused with prediction parity as defined by Chouldechova (2017).
- Cost ratio parity is given by $[[\mathbb{P}(\hat{Y} = 1|Y = 0, a)/\mathbb{P}(\hat{Y} = 0|Y = 1, a)] = [\mathbb{P}(\hat{Y} = 1|Y = 0, a^*)/\mathbb{P}(\hat{Y} = 0|Y = 1, a^*)]]$ and requires that the ratio of false positives to false negatives (or the inverse) is the same for similarly situated men and women. This is important because the cost ratio indicates the rate at which a risk algorithm trades false positives against false negatives. In practice, the cost ratio is a key tuning parameter used to respond to the perceived relative costs of different kinds of classification errors (Berk 2018). For example, many stakeholders will claim that it is much worse to incorrectly classify an offender as low risk than to incorrectly classify an offender as high risk. The former is a strong candidate for a postrelease arrest. The latter could be needlessly incarcerated. Both mistakes have costs, but the costs are unlikely to be seen as equal. Such costs are rarely monetized because many of the most important consequences, such as psychological trauma, are difficult to quantify. What matters is the unit-free ratio of costs.

Several closely related fairness definitions have been employed or proposed. They too can be computed from a confusion table or test data more generally, but there is no agreement about whether they should be preferred to those just discussed or even if they usefully supplement those definitions. We do not address them here, and useful discussion can be found elsewhere (Baer et al. 2020, Mitchell et al. 2021).

Some recent fairness developments abandon conventional parity-based definitions altogether. For example, Diana et al. (2021) propose minimizing maximum algorithmic error across all protected groups, subject to the constraint that the protected group with the most algorithmic error has a minimax error smaller than some predetermined constant. On the average, all groups gain improved protection against the worst outcome, and no single protected group must bear an unacceptable protection burden. Algorithmic error can be defined in several ways, including classification error or forecasting error, much as one might compute from a confusion table. The proposal is technically sophisticated and well-executed, but it remains to be seen how it will be received in jurisprudential circles and among stakeholders with boots on the ground. Minimax loss for protected groups is neither a self-evident truth nor an easily inferred extension of any civil rights statutes or precedents. It is also not clear why minimax loss is a desirable criterion when, as a matter of past research or credible subject-matter theory, that loss will be almost certainly overly pessimistic.

Romano et al. (2019) build their interesting fairness approach around equal coverage of prediction intervals across all protected groups. Their approach offers distribution-free coverage with finite sample guarantees for exchangeable data, building on recent developments in conformal inference (Lei et al. 2018) and extensions to conformal quantile regression (Romano et al. 2019). They offer a form of uncertainty fairness. But here, too, no jurisprudence justification is evident.

Other recent additions to the fairness definition compilation stem from causal inference reasoning commonly represented in directed acyclic graphs (DAGs) (Kussner et al. 2018, Nabi & Shpitser 2018, Wu et al. 2019, Baer et al. 2020, Oneto & Chiappa 2020, Mitchell et al. 2021). Using conventional causal inference formulations, a key goal is to introduce counterfactual reasoning that can be used to identify the causal paths responsible for unfairness. For example, race may directly affect a hiring decision and/or indirectly affect a hiring decision through educational attainment. Although reasoning in this manner can help clarify the sources of unfairness, all of the usual obstacles surface when applied to a real setting with the data likely to be available. The user has an implicit structural equation model that must be correctly specified in principle and in practice. These are well known to be daunting requirements (Freedman 2006), particularly for identification used in causal path analysis (Avin et al. 2005). More discussion of models is beyond the scope of this article (see Section 2).[1]

Finally, definitions of fairness sometimes are stretched in a manner that dilutes or even eliminates any moral content. For example, Pan et al. (2021) consider problems caused by an unbalanced or skewed response variable. They use the term minority when referring to cases with underrepresented outcome categories or rare numeric values. Such observations can experience algorithmic unfairness because their Y-values are more likely to be estimated poorly compared with the Y-values of other cases. Thus, random forests would be unlikely to partition on rare observations when each tree in the forest is grown and probably would return less accurate fitted values for those observations. But there is no mention of protected groups.

5. FAIRNESS COMPLICATIONS

Although constructing a proper confusion table with test data is easy to do, arriving at valid interpretations of fairness can be challenging. There are difficulties in how fairness is empirically evaluated, but also problems preceding fairness assessments that make them difficult to properly execute.

[1]Some treat a model's computational machinery by itself as an algorithm and interpret that machinery in the same explanation-free manner. This is sometime done in computer science, for instance, with linear regression (Mohri et al. 2018); $(X'X)^{-1}X'Y$ is the algorithm. What matters then is solely prediction computed in the usual linear fashion.

5.1. Classifier Reliability

Algorithmic classifiers usually compute a score or probability before an outcome class is assigned. Sometimes that score or probability indicates that the classifier is able to make clear distinctions between outcome classes, and sometimes it does not. Fitted probabilities of 0.51 compared with 0.49 for a binary outcome convey that one outcome class is predicted to be nearly as likely as the other. Fitted probabilities of 0.91 compared with 0.09 convey that the first predicted outcome is far more likely than the other. Yet both pairs of probabilities can classify a case identically if the same outcome class has the larger probability in both comparisons. In other words, the reliability of an assigned outcome class matters. Classes assigned with low reliabilities could have easily been different because of randomness in how units are realized, randomness built into the algorithm (e.g., in random forests), random splitting of a data set into training and test data, and small changes in the covariates used. One would prefer not to make a hiring decision, for example, on little more than the flip of a fair coin.

5.2. Measurement Error

There are often serious measurement problems in the data used to train and test algorithms. Worse, measurement error typically is not random and often is associated with protected group membership. Such concerns have been raised perhaps most vocally about criminal justice decisions because so much of the data available are a product of data collection for administrative purposes by criminal justice agencies. For example, arrests and convictions conflate crimes actually committed with the practices of police, prosecutors, and judges, all of whom have had their impartiality questioned. Making more transparent assumptions about measurement quality can surely help (Cooper & Abrams 2012), but because measurement error can be ubiquitous and systematic, remediation of various kinds must be considered, even if the available repairs are not fully satisfactory. Measurement concerns are raised often in the pages ahead. Perhaps the most important controversy is that measures assumed to yield legitimate variables for the setting in question actually may fold in illegitimate processes, often related to race and gender. The criticisms of standardized tests meant to inform college admissions decisions are one illustration.

5.3. The Need for an Outcome Standard

One must have an answer to the question, fairness with respect to what outcome standard? In the Berkeley example, should the overall admission rate be the overall male admission rate, the overall female admission rate, or some other admission rate? One possibility would be to link a common admission rate to an ideal incoming class size.

If an outcome standard is not specified as part of a risk algorithm, the algorithm can make its own, often inappropriate, determination. In the extreme, everyone can be made equally worse off. This is a long recognized problem in efforts to remedy race and gender discrimination in sentencing (Fisher & Kadane 1983), which seems to have escaped notice in most work on algorithmic fairness.

A very simple and instructive algorithm in which outcome standards are established is the sentencing guidelines promulgated by the US Sentencing Commission (**https://www.ussc.gov/**). Convicted individuals with the same offense characteristics, often ranked for seriousness (Rossi et al. 1974), and the same offender characteristics (e.g., number of prior felony arrests) are recommended for the same penalty, such as 5 to 10 years in prison. The target sentence range is a product of earlier deliberations by members of the Commission after substantial public input. In the end, judges are provided a very clear answer to the question: fairness with respect to what punishment?

5.4. The Meaning of "Similarly Situated"

The meaning and operationalizing of "similarly situated" have been treated as some variant on (*a*) employing an auditor's intuition (Maity et al. 2021); (*b*) avoiding the use of legally protected class attributes such race, ethnicity, religion, and gender; or (*c*) adjusting for racial confounding with some technical fix. The first is essentially a placeholder that acknowledges a potential problem but actually solves nothing. The second is well-known to be inadequate because many included covariates will be associated with protected class attributes (Mitchell et al. 2021). Protected class membership inadvertently can be incorporated into the analysis. The third automatically fails when it ignores a range of normative and legal factors affecting which covariates are legitimate and which are not. These complications shape how confounding should be handled.

For example, under title VII of the US Civil Rights Act of 1964 (Pub. L. 88-352 §625), a person's sex, religion, and national origin can be "bona fide occupational qualifications." Sex, religion, and/or national origin are legally acceptable attributes if necessary to successfully undertake a job that is a normal activity for a business or other enterprise. Thus, in *Hosanna-Tabor Evangelical Lutheran Church and School v. Equal Employment Opportunity Commission* (2010), the US Supreme Court ruled that a person's religion can be a bona fide occupational qualification for teaching in private religious schools. In that setting, religion was no different from other employment qualifications such as teaching experience. The ruling did not apply to public schools even for religious course content.

Faulty formulations of "similarly situated" often highlight insufficient understanding of the differences between disparities and unfairness. For instance, men are enormously overrepresented in prisons across the United States. Clearly, there is a very large gender disparity. But throughout recorded history, men have been responsible for the vast majority of violent crimes. Can one credibly argue that the overrepresentation of male inmates is primarily a product of gender unfairness in the criminal justice system? If an algorithm routinely and accurately forecasts that men are a greater risk to public safety than women, is that algorithm being unfair? And gender disparities are but one illustration. The differences between disparities and unfairness need to be determined in each application setting before there can be agreement on what constitutes similarly situated cases.

5.5. Challenging Trade-Offs

Except in highly stylized or improbable settings, there are often provable trade-offs between different kinds of fairness and between fairness and accuracy (Kleinberg et al. 2017). As a result, there commonly can be no technical fix such that in practice stakeholders can have it all. Difficult choices must be made about which kinds of fairness are most important and how their various trade-offs will be determined. In real applications, these choices should not be made for mathematical convenience. Difficult and often contentious discussions among stakeholders are required before acceptable compromises between competing interests and perspectives can be reached (Berk & Elzarka 2020).

5.6. Individual Fairness

Notwithstanding our emphasis on fairness across groups, there is a small but insightful literature on individual fairness (Dwork et al. 2011, Gillen et al. 2018, Maity et al. 2021). Similarly situated individuals must be treated similarly. A major critique of group fairness is that individual fairness can be sacrificed to the greater good, although there is some debate about whether that represents a fundamental incompatibility between the two (Binns 2019).

Perhaps the most challenging problem for individual fairness is to define and implement a numeric fairness scale capturing an acceptable form of normative fairness that is tractable. In

somewhat stylized treatments (Bolukbasi et al. 2016, Madaan et al. 2018), and/or with strong modeling assumptions (Mukherjee et al. 2020), this can be accomplished. Currently, however, it would be an enormous challenge to apply such work to the significant societal issues of the day (e.g., health care, employment, climate change, public safety) precisely because these are the very problems for which there are typically strong disagreements about how to define and operationalize fairness. For example, how unfair would it be to bar a transgender woman from competing as a woman in intercollegiate sports (Binder 2022)?

5.7. Missing Follow-Up Data

Standard practice has risk algorithms trained and evaluated with data readily available. Sometimes this is a matter of convenience, and sometimes there is effectively no choice. In either case, the data capture existing practices, whereas the algorithm represents practices that have not (yet) been implemented. In particular, it is those historical practices that are to be intentionally altered so that the future is not like the past. This often creates a major problem with missing data because the future has not yet happened.

For simplicity, consider fairness for the false positive rate. For a given commercial bank, an observed false positive rate might be the fraction of business establishments in the past that defaulted on a small business loan. Presumably, a loan was provided because bank officials believed such businesses were a promising investment. One might compare the false positive rate for Black-owned businesses to the false positive rate for White-owned businesses as a special case of classification parity.

Suppose that the bank officials asked their information technology (IT) department to develop a risk algorithm that reduced the bank's financial exposure but, as a byproduct, made their lending decisions more fair. Necessarily, that algorithm would be trained and tested on the historical data. One likely estimate of improved fairness would be the original false positive fairness compared with the false positive fairness under the risk algorithm, both computed with the existing test data. And in turn, that comparison might be important in deciding whether to implement the algorithm to inform future lending decisions. Was there evidence from the historical data that the risk algorithm would improve the false positive fairness if the algorithm were put in practice?

The whole purpose of such an algorithm would be to reform the process by which lending decisions were made and, consequently, to change the mix of small businesses receiving loans. A likely repercussion is that the bankruptcy base rates for Black-owned and White-owned businesses would change, perhaps in different directions. More generally, all results from the original confusion tables, including measures of fairness and accuracy, would likely differ. But data on those changes would not exist when a decision was needed on whether, going forward, to apply the algorithm to inform real decisions. Recall that the Berkeley admission example faced similar challenges.

The problem can be framed more broadly as making credible inferences about counterfactual phenomena when a risk algorithm is viewed as an intervention in the status quo, and only preintervention data are readily accessible in a timely fashion. Proper postintervention data are rarely available when an implementation decision needs to be made, but these are the very data required to properly evaluate a risk algorithm's future accuracy and fairness. We consider the issues far more deeply in Section 8 after further foundational material is provided.

6. ADJUSTMENTS TOWARD FAIRNESS

There is wide agreement that the usual data on which risk algorithms are trained and tested can contain variables badly tainted by illegitimate influences. There is also widespread agreement that simply discarding illegitimate covariates will not suffice because the tainted covariates typically

are associated with covariates taken to be legitimate (Baer et al. 2020, Mitchell et al. 2021). There is nothing close to a consensus on how best to remove the objectionable content, but there are many creative proposals for working with training and test data.

6.1. Preprocessing Proposals

Guided by knowledge of the subject matter and measurement procedures, one simple form of preprocessing is transforming suspect covariates. For example, many argue that an offender's prior record builds in inappropriate race and gender differences because of police practices or biases. The number of prior arrests arguably is, for many offenders, inaccurately inflated. For such offenders, a square root transformation applied to a prior arrest count will pull in a long right distribution's tail (Berk & Elzarka 2020). Alternatively, for race or gender comparisons, one might use a function of the quantiles, rather than the raw count. For instance, a male offender at the 90th percentile for men could be given the same prior record score as a female offender at the 90th percentile for women, even if the male offender had many more prior arrests. Such methods are judgement calls in the sense is that there is no objective function being optimized.

Far more statistically principled procedures are readily available. For example, Kamiran & Calders (2012) focus on the special case of problematic covariates such as gender. Suppose the outcome is whether a job candidate is hired or not, and there is already evidence that women are underrepresented among those hired, compared with similarly situated men. The key idea is to apply an algorithm to the training data that changes some outcome class labels from not hired to hired for women and some outcome class labels from hired to not hired for men so that, subsequently, an acceptable trade-off between fairness and accuracy is achieved. The numbers of men and women affected should be the same so that the marginal distribution of the outcome is not altered. Cases chosen for label swapping are those least strongly associated with the hiring decision, such that accuracy is compromised as little as possible, and the number of labels swapped is a tuning parameter. The goals of label swapping also can be achieved by reweighting or disproportional stratified sampling.

Calmon et al. (2017) take a rather different approach, proposing convex optimization that changes the composition of the data itself. The data on hand are used to approximate the generative joint probability distribution. Then, observations from the training and test data are randomly replaced with draws from that distribution in a manner that (*a*) limits problematic associations between the response variable and protected class membership, (*b*) prevents distortions in individual observations larger than some specified constant, and (*c*) approximately maintains the data's joint probability distribution. The new data set is then used for training.

6.2. In-Processing Proposals

Procedures to improve fairness while an algorithm is trained are probably the most common ways to remove objectionable feature content. Available methods can be very sophisticated technically but are often overly stylized for mathematical convenience. They also typically introduce an accuracy-fairness trade-off.

For example, Corbett-Davies et al. (2017) propose a form of constrained optimization that aims to maximize public safety under a specified fairness constraint. A key mechanism is that risk thresholds on algorithmic scores or probabilities that determine fitted classes can differ by protected class memberships. Thus, if for Black offenders, prediction parity is violated by more common incarceration sentences, their classification threshold is increased relative to other offenders such that their incarcerations become less common. With the same intent, Feldman et al. (2015) advocate using within-group percentiles. Black offenders whose risk score or probability

places them at, say, the 75th percentile for their risk distribution would be given the same sentence as White offenders at the 75th percentile for their risk distribution.

Berk et al. (2021) also target uncertainty fairness using conformal inference to address prediction parity. Suppose again that there are Black offenders and White offenders characterized as less advantaged and more advantaged, respectively. A classifier is trained only on White offenders, which precludes baking in any racial differences. Then, the joint covariate distribution of Black offenders is conveyed, using optimal transport (Peyré & Cuturi 2019), to the joint covariate distribution of the White offenders, which equalizes the two joint predictor distributions. In these two steps, Black offenders are treated by an algorithm as similarly situated White offenders. Finally, conformal inference is applied to obtain comparable prediction set distributions.

6.3. Postprocessing Proposals

Sometimes there are concerns about fairness in algorithmic results, but there is no access to the original training and test data. All one has is the algorithmic output. For example, there is no opportunity to reconsider whether appropriate distinctions had been made between legitimate and illegitimate covariates. Nevertheless, fairness can be improved by postprocessing the output. For example, Kim et al. (2019) use a form of boosting, after a fairness audit informed by a small validation data set, to reweight the output such that differences in classification accuracy across identifiable protected groups can be eliminated.

Perhaps the most well-known approach is the one taken by Hardt et al. (2016), in which fairness requires equalized odds. For a binary Y coded as 1 or 0, where 1 is the more favorable outcome, equalized odds is defined as

$$\mathbb{P}(\hat{Y} = 1|Y = y, X, a) = \mathbb{P}(\hat{Y} = 1|Y = y, X, a^*), \quad y \in \{0, 1\}.$$

In words, the true positive rate is the same for protected class a and protected class a^* and the false positive rate is the same for protected class a and protected class a^*. Note that for postprocessing, the classifier is only available as a black box. Also, when forecasts are needed, the true Y is not known. Formally, one applies linear programming to construct a new $\hat{Y}$ distribution $\tilde{Y}$ that minimizes an expected loss $\mathbb{E}[l(\tilde{Y}, Y)]$ subject to constraints of equalized odds and the values of four parameters, each ranging from 0 to 1. These parameters are the conditional probabilities for true positives and negatives and false positives and negatives. Their values are determined to provide a linear programming solution. Basically, some outcome labels are being judiciously flipped.

Hardt et al. (2016) are highly critical of using prediction parity as a measure of fairness. Le Gouic et al. (2020) apparently disagree and make prediction parity their measure of fairness. They focus on the trade-off between accuracy and fairness for a class of regression functions and provide a way to find the regression function in which the trade-off is most favorable. Accuracy is defined through squared Wasserstein distance, and the objective is to minimize excess risk, which is the expected difference between the squared Wasserstein distance for a regression with no fairness adjustments and a regression with fairness adjustments. A fairness adjustment is successful if prediction parity is substantially improved with little damage to accuracy. The adjustment procedure uses optimal transport to couple the distribution of the unfair regression fitted values toward the distribution of fair fitted values to arrive at the optimal trade-off.

For example, thinking again about hiring decisions, suppose there are three protected group classes: White, Black, and Latinx. There are optimal fitted values from a regression in which the protected group classes are not included covariates. These fitted values are indirectly associated

with the protected group classes. Fitted values are computed separately for each protected group class, and optimal transport is used to couple each set of fitted values, in turn, to fitted values of the other two groups. In effect, one is making the distribution of fitted values for each group comparable to fitted values of the other groups. So, for example, the distribution of fitted values for Black applicants is made comparable to the fitted value distribution for job applicants who are White or Latinx. The process is repeated for each protected class. The three sets of transported values are then combined as a weighted average for the fair fitted values. Chzhen et al. (2020) arrive at very similar results.

6.4. Fairness Adjustment Conclusions

There is no agreement about how best to define fairness or about the best methods to achieve it. The available adjustments toward fairness typically consider one particular fairness definition, or a small subset of definitions, implicitly agreeing with Kleinberg et al. (2017) that in practice you cannot have it all. Further complicating comparisons across methods is substantial variation in formulations and assumptions even when fairness is defined in the same manner. To that we add that because the foundational concerns raised earlier are largely ignored (see, again, Section 5), there can be considerable conceptual ambiguity and telling oversights even when substantial effort is otherwise invested in clarity.

7. FAIRNESS UNCERTAINTY

The forecasting goal motivating risk algorithms necessarily introduces concerns about uncertainty, which in many settings is difficult to characterize properly. For algorithmic fairness, one begins by requiring independent and identically distributed (IID) or exchangeable case realizations. Care must be taken to avoid risk algorithms that can compromise this requirement, or proper remedies, such as estimation with a holdout sample, must be introduced. Model selection and data snooping more generally are frequent culprits (Kuchibhotla et al. 2021).

Uncertainty arises most explicitly at three different places in the development and implementation of risk algorithms: for (*a*) aggregate classification accuracy and aggregate forecasting accuracy, (*b*) fairness determinations, and (*c*) the label(s) forecasted for new cases (Berk et al. 2021). For aggregate algorithmic performance assessments computed from test data, conventional X–Y resampling methods can be applied; the training data are treated as fixed. Uncertainty for both accuracy estimates and fairness estimates are readily obtained. For example, several hundred X–Y bootstrap samples from the test data will generally provide a good empirical approximation of the sampling distribution for the false negative rate from a given confusion table. The resampling approach can be extended to classification parity and/or forecasting accuracy parity for two confusion tables constructed for two different protected groups; two proportions are estimated from two different IID samples to consider uncertainty for fairness estimates. In short, any of the usual summary statistics from such confusion tables are fair game.

More novel are uncertainty estimates attached to forecasted labels for individual cases using nested conformal prediction sets (Kuchibhotla & Berk 2020). With IID or exchangeable training and test data, a data analyst can obtain, for any forecasted best prediction set, valid finite sample coverage at a selected probability. For example, the single forecasted label for a particular mortgage applicant might be default, which is guaranteed to be the true label at, say, a probability of 0.95. Over a large number of cases for which such forecasts are made, claims that a prediction set contains the true outcome will be correct about 95% of the time. The same reasoning can, in principle, be applied to subsets of cases defined by their covariate values (e.g., men 25 years of age with a credit score below 600), although the inferential guarantees are now asymptotic.

When the classifier cannot make definitive outcome distinctions for a particular case, the prediction set can include more than one label. An example would be the label for a mortgage default and the label for no mortgage default. In finite samples, one still gets a valid coverage probability.

Over many forecasts, one usually can also evaluate prediction parity for a given coverage probability (Berk et al. 2021). Suppose a mortgage default is coded as 1, and no mortgage default is coded as 0. The coverage is set at 0.95. There is prediction parity if, over many cases, the probabilities across protected groups for each possible prediction set are effectively the same. For example, prediction set compositions with their probabilities of occurrence for one protected group might be $p(\emptyset) = 0.01$, $p(\{0\}) = 0.20$, $p(\{1\}) = 0.44$, and $p(\{0, 1\}) = 0.35$, where $\emptyset$ is an empty set (i.e., cases that are so atypical that prediction sets cannot be constructed). The distribution for other protected groups should be approximately the same.

8. A FRAMEWORK FOR RISK ALGORITHMS AS INTERVENTIONS

With few exceptions (Imai et al. 2022), the computer science and statistics literature concentrates on building effective risk algorithms with historical training and test data. We have already indicated that from a practical perspective, an essential step is missing. Thinking back to the Berkeley admissions illustration, if gender unfairness were empirically verified, changes in graduate admission procedures would likely have been seriously considered. These days, reforms might be embodied in a risk algorithm whose primary objective would be accurately forecast an applicant's academic performance. Also, the algorithm would better achieve one or more kinds fairness for protected groups.

However, at the time a decision would be needed about whether to deploy such an algorithm, there would be no data available to estimate properly how that algorithm would perform if used subsequently to inform real admissions decisions. To take a simple example, some students in the test data who were not admitted would be admitted if their admissions decisions were algorithmically informed. One could obtain their forecasted Berkeley academic performance using the trained algorithm. But, there would no Berkeley classroom performance information for these applicants and, therefore, no way to evaluate the accuracy of those forecasts or their contributions to classification parity or forecasting accuracy parity. Schökopf (2019, p. 2) quotes Shannon (1959) as saying, "...we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it." This statement has several implications for this discussion.

As in Berk et al. (2021), we call estimated accuracy internal when it is computed from the usual test data. We call the accuracy we most care about in practice external because it must be estimated from data beyond the data available when a risk algorithm is trained and tested; it is external to the data one has when a decision needs to be made about whether to deploy a risk algorithm. The same reasoning is applied to fairness. Internal fairness is estimated from test data. External fairness, which can be far more instructive in practice, is estimated from data unavailable when a decision must be made about deploying a risk algorithm.

These and related distinctions are usefully placed in a larger context. A risk algorithm is an intervention in the status quo. Estimands of interest center on disrupted business as usual. Because risk algorithms impose a threshold on a classifier's output, the risk algorithm can be viewed as a defining feature of a quasi-experimental, regression discontinuity design (Imbens & Lemieux 2008). Regression discontinuity designs have been employed effectively since the 1960s (Campbell & Stanley 1963). When the design is properly implemented, there can be excellent prospects for valid causal conclusions about algorithmic performance in practice, not just from extrapolated training and test data. A major requirement, however, is that the risk algorithm be implemented on a provisional basis that allows for revisions, updates as the setting changes, and even the option

to abandon the risk algorithm if it fails to perform adequately, all based on postintervention data. This might be a hard sell in certain settings but seems rather sensible as a general principle for any programmatic or policy intervention and is in the same spirit of recent work by others (Shi et al. 2021, Imai et al. 2022).

Consider a variant on the small business loan example. Suppose the IT staff of a given bank has developed a risk algorithm that forecasts whether a small business applying for a loan will declare bankruptcy within 2 years after applying for a business loan. Declaring bankruptcy is coded 1, and no declaration of bankruptcy is coded 0. This information is available from bank records and usually is in the public domain more generally.

The bank wants to improve its ability to pick winners: firms that will repay the loan with interest. But fairness is also a concern. Black-owned small businesses seeking a loan should be treated no differently by the algorithm from similarly situated White-owned small businesses. For the moment, assume that the bank will follow the algorithm's recommendations.

Recall that for classifiers, $\hat{p}_i = \Psi(V = v_i)$ can be a probability on which one imposes a threshold such as $c = 0.50$, so that a study unit i has its loan application denied if $\hat{p}_i > 0.50$ and otherwise receives a loan. Cases for which $\hat{p}_i$ is very near c effectively have the same risk probability, yet those above the threshold receive one intervention (i.e., no business loan), while those at or below the threshold receive the other intervention (i.e., a business loan). The difference in the proportion of firms that declare bankruptcy near to, but on either side of the c is an estimate of the average treatment effect of the risk algorithm on bankruptcy. Formal details and assumptions for the regression discontinuity design are provided by Imbens & Lemieux (2008). There are also extensions that can capitalize on the full range of $\hat{p}$ values, not just those very near c, which in principle can increase statistical efficiency (Berk & DeLeeuw 1999). A key requirement is that the (perhaps nonlinear) functional form of $\hat{p}$ is the same on both sides of the threshold, save for the intervention.

The regression discontinuity design can provide valid causal conclusions about the impact of an algorithmic intervention. Is the algorithm picking winners well and in a fair manner? This information will be readily available from a conventional confusion table but constructed from postintervention data. The uncertainty tools described earlier can also be applied.

The design is also easily extended to applications in which an algorithmic recommendation is not definitive. For example, the recommendation might be among several inputs to a bank employee, who ultimately makes the lending decision. One now has a fuzzy regression discontinuity design that is also easily analyzed because one knows when the algorithmic recommendation is followed and when it is not. As before, details and assumptions are provided by Imbens & Lemieux (2008).

At the same time, the limits of an algorithm's reach must not be forgotten. In this example, any improvements in fairness end once a loan decision is made. The business environment faced by Black-owned businesses or White-owned businesses is almost certainly unchanged by the risk algorithm itself. The algorithm cannot be held responsible for the fairness of business institutions and markets.

The regression discontinuity design, like any design from which causal inferences are sought, can be badly compromised by missing or incomplete postintervention data on the outcome, just as in the Berkeley admissions example. Suppose that the bank can only learn if a firm declares bankruptcy when that firm was among those to which the bank made loans. There are no follow-up data on the firms whose loan application was denied. Ideally, there will be access to supplementary data with the requisite information. For example, the bank might begin requiring that all loan applicants, whether successful or not, provide their year-end reports or even tax filings.

There can be a range of fallback positions if such data are produced in other settings. For example, bankruptcy information is available from the US Trustee Program of the US Department

of Justice, although its coverage is not complete and gaining access might be difficult. The general point, however, is that the populations represented in new data sources might differ from the population of immediate interest. But these kinds of complications may be less problematic than the complications from causal inferences using solely the historical training and test data.

9. FUTURE DIRECTIONS

For all of the complications addressed to this point, there are ways significant progress can be made. What follows are some suggestions. Technical challenges, often of particular interest to statisticians, are commonly embedded within broader concerns. We start with the broader concerns.

9.1. Meta-Issues

Future work on fair risk algorithms would benefit from a more careful consideration of actual applications as a foundation for formal statistical advances. Such an approach might consider the following issues.

- Academic discourse and real applications: For academic purposes, there is enormous freedom to tailor the questions and concepts for mathematical convenience. When applications are driving the enterprise, key questions and concepts must originate from a host of practical considerations, even if the resulting mathematical challenges can range from trivial to intractable. The realpolitik matters deeply, even if it can be messy. But differences in goals and methods between theoretical and applied work are not license for one approach to ignore the other. Vigorous hand waving to link the two does not count.
- Settings matter: The settings in which fair risk algorithms can properly be used are heterogeneous, and each can differently shape algorithm development, adoption, and use. For example, gender may be a protected class under some circumstances but not others. Even for theoretical work, the settings envisioned should be explained.
- Actual use matters: When there is interest in making fair risk algorithms useful, their developers should consider how the risk forecasts will be employed. For example, risk algorithms often are assumed to fully determine some decision; humans are not in the loop. Yet, in criminal justice settings, say, it is very difficult to find decisions for which that is true. Rather, a risk algorithm is intended help inform decisions, although in practice algorithmic output can be completely ignored or misconstrued (Stevenson & Doleac 2018, Imai et al. 2022). In other governmental settings, as well, humans more typically make the decision even if algorithms provide useful information (Coglianese & Lehr 2018). The operational goal is providing helpful direction, not replacing humans. These two different decision processes have different implications for how the algorithmic results are conveyed and the kinds of transparency provided. The ways in which accuracy and fairness are evaluated likely will also differ.
- Risk algorithms have limited reach: Complaints about risk algorithms often fold in, at least implicitly, decisions and actions for which algorithms are not responsible. Misleading conclusions can follow and can lead to misdirected reform efforts. Once again, understanding of the setting is vital, and media coverage is not necessarily an accurate resource; a useful peer reviewed literature and genuine expertise are often available.
- Outcome standards: There are important conceptual precursors to definitions of fairness. Among the most important is the outcome standard adopted. For example, any hospital will have an upper limit per day to the number of patients who can receive, say, proton therapy for cancer. That upper limit likely will be the outcome standard for a fair algorithm meant to

inform decisions about which patients are offered treatment and becomes, de facto, a form of health care rationing (Bognar & Hirose 2014). If that upper limit is not built into the algorithm, whether fairness would be achieved may be moot. And if the hospital's capacity to offer proton therapy grows, the algorithm will probably require a significant revision. The threshold determining which patients to treat might be reduced.

- Valid comparisons: Fairness definitions are fundamentally comparative and ideally require that the study units to be compared are similarly situated on legitimate covariates. Legitimacy is not a statistical matter because it depends on social norms, regulations, statutes, and prevailing legal precedents. But unless these factors guide an algorithm's development, the algorithm risks being irrelevant.
- Measurement matters: Many unnecessary mistakes could be avoided with better understanding of how the variables used to develop fair algorithms are measured. For example, an argument is sometimes made that for criminal justice applications, prior record should be measured by convictions rather than arrests. Arrests are seen as more prone to misrepresenting the real crimes and are vulnerable to the racial bias of some police officers. Yet, the vast majority of convictions result from plea bargains that are shaped by local prosecutorial practices, race and gender stereotyping, pressures from media coverage and upcoming elections, availability of witnesses, bullying of defendants, all manner of strategic maneuvering, and procedural convenience (Bielen & Grajzi 2021). There is no reason to think that the crimes for which an offender is convicted are closer to the true crimes than the crimes for which an offender is arrested, and making that claim can fundamentally compromise an algorithm's credibility very early in its development.
- How the data were generated matters: It is widely recognized that data used to develop any risk algorithm should be realized from the joint probability distribution that also represents the population or process of interest. This also applies to the unlabeled cases for which forecasts are needed. But convenience can trump prudence, and often it may be difficult to undertake valid statistical inference or to know the settings to which a particular risk algorithm can be properly generalized. These inferential challenges must be addressed and thoughtful caveats provided as needed.
- Preintervention and postintervention data: It seems that too few of those who develop risk algorithms fully appreciate the limitations of preintervention data. The intent of deploying a fair risk algorithm is to alter the status quo. The preintervention training and test data are, therefore, questionable proxies for the algorithm's performance. The limitations of preintervention data should always be made clear, and efforts should be made to collect good postintervention data with a strong research design. Even randomized experiments are possible (Imai et al. 2022), but a regression discontinuity design can flow naturally from the way risk algorithms can be deployed. We suspect, however, that the major obstacle will be objections to implementing risk algorithms on a provisional basis that are motivated in part to collect outcome data. There may well be promising ways to respond over the long term with biomedical randomized experiments.

9.2. Technical Challenges

There are also many technical challenges whose details depend on how the meta-issues are handled.

- Causal inference for algorithmic interventions: Causal inference has surfaced in several of our earlier discussions, often informed by a substantial literature. We initially introduced the need for causal inference while setting the stage with the Berkeley admissions

illustration. The use of DAGs in definitions of fairness is a more weighty example. Perhaps less widely recognized is the necessary role of causal inference for evaluating the impact of algorithmic interventions. Beyond the potential use of randomized experiments and regression discontinuity designs, there have been a number of recent developments in program evaluation that might be useful. For example, some states are now discussing whether to make formal risk assessments mandatory for sentencing. Were such procedures adopted, a credible evaluation might be undertaken with synthetic cohort methods using states as the observational units (Robbins et al. 2017). All of these tools themselves can be improved, but there are significant execution challenges because provisional implementations needed for postintervention data may be a hard sell.

- Systematically missing outcome data: In many important applications, one may not get to see the outcome for a systematic subset of cases. Going back to the Berkeley example, one cannot observe the Berkeley academic performance of graduate school applicants who were not admitted. Fairness for such individuals cannot be empirically examined. Earlier, we provided some initial ideas that might help, but surely there is much more that can be done with the rich technical literature on missing data, domain adaptation, and transfer learning (Pan & Yang 2009).
- Risk algorithm generalizability: An ongoing challenge is applying risk algorithms developed in one setting to one or more other settings. For example, can a risk algorithm developed for a given hospital's patient readmissions be used effectively in other hospitals? Generalizability is an important and active area in machine learning more broadly that includes such topics as covariate shifts (Tibshirani et al. 2020), transfer learning (Zhuang et al. 2020), anchor regression (Rothenhäusler et al. 2021), and domain adaptation (Chen & Bühlmann 2021). Without substantial progress, each setting will require its own hand-tailored risk algorithm. Clearly, this is impractical.
- Classifier improvements: Most of the ongoing research on classifiers in statistics and computer science is relevant. Examples of areas in which there are promising developments include the double descent phenomenon (Nakkiran et al. 2019), causal inference with machine learning (Schökopf 2019), and valid statistical inference for machine learning (Kuchibhotla et al. 2021). More such work is needed.
- Adjustments for confounders with protected groups: Many options already exist that can be used with preprocessing, in-processing, and postprocessing (Berk et al. 2018). There has been little effort to systematically address the trade-offs between these different methods. How is a practitioner to choose between them?
- Documenting trade-offs: There is already substantial work trying to document the trade-offs between different kinds of fairness (Kleinberg et al. 2017) and between fairness and accuracy (Cooper & Abrams 2012). The goal is to find trade-off sweet spots where fairness is increased substantially with small accuracy losses. This work can be enriched with higher-dimensional trade-offs, such as between prediction parity, classification parity, and forecasting accuracy (Rodolfa et al. 2021). Performance and acceptance of fair risk algorithms could be substantially improved with such knowledge.
- Better software: The references in this review depend heavily on posted papers that have not yet been peer reviewed. It is not surprising that much of the available software is best thought of as prototypes not easily employed by individuals who are not insiders. The movement from theory to practice will be accelerated by software that is more widely accessible and less dependent on insider knowledge. Shareware is preferable so that those who wish can fully learn how a procedure works. Also, it can be difficult to evaluate claims made by purveyors of proprietary code.

10. CONCLUSIONS

Few would argue with the ideal of fair risk algorithms. Fairness holds the moral high ground and introduces many technical problems of great interest to statisticians and computer scientists. On closer inspection, however, there are many kinds of fairness that create difficult trade-offs and strongly defended normative claims. The trade-offs present significant technical challenges, some of which may turn out to be intractable. The normative claims vary across stakeholders and settings, and are made even more nettlesome by legal, economic and administrative overlays. It is not surprising, therefore, that the relevant literature, which is extremely rich in ideas, is badly fragmented. But there is also self-inflicted damage. More progress might be made with better engagement between technical work and practical work. The disconnect means that theoretical formulations can be tangential to fairness in practice. The moral high ground looks more like pandering. Meanwhile, work on real applications can fail to capitalize on genuine technical advances or, worse, encourage technically ill-informed fairness policies. Finally, it is, in our view, critical to recognize that if statistical work on fair algorithms is to be properly evaluated, the algorithms should be treated as interventions that require postintervention data or very good proxies for those data.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Avin C, Shpitser Id, Pearl J. 2005. Identifiability of path-specific effects. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pp. 357–63. N.p.: IJCAI

Athey S, Imbens G. 2019. Machine learning methods that economists should know about. *Annu. Rev. Econ.* 11:685–725

Baer BR, Gilbert DE, Wells MT. 2020. Fairness criteria through the lens of directed acyclic graphs: a statistical modeling perspective. In *The Oxford Handbook of Ethics of AI*, ed. MD Dubber, F Pasquale, S Das, pp. 493–520. Oxford, UK: Oxford Univ. Press

Becker GS. 1996. *Accounting for Tastes*. Cambridge, MA: Harvard Univ. Press

Ben-Michael E, Greiner DJ, Imai K, Jiang Z. 2022. Safe policy learning through extrapolation: application to pre-trial risk assessment. arXiv:2109.11679v3 [stat.ML]

Berk RA. 2017. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *J. Exp. Criminol.* 13(2):633–55

Berk RA. 2018. *Machine Learning Forecasts of Risk in Criminal Justice Settings*. New York: Springer

Berk RA. 2020. *Statistical Learning from a Regression Perspective*. New York: Springer. 3rd ed.

Berk RA, DeLeeuw J. 1999. An evaluation of California's inmate classification system using a generalized regression discontinuity design. *J. Am. Stat. Assoc.* 94(448):1045–52

Berk RA, Elzarka A. 2020. Almost politically acceptable criminal justice risk assessment. *Criminol. Public Policy* 19(4):1231–57

Berk RA, Freedman DA. 2003. Statistical assumptions as empirical commitments. In *Punishment and Social Control*, ed. TG Blomberg, S Cohen, pp. 234–58. Piscataway, NJ: Aldine de Gruyter. 2nd ed.

Berk RA, Heirdari H, Jabbari S, Kearns M, Roth A. 2018. Fairness in criminal justice risk assessments: the state of the art. *Sociol. Methods Res.* 50(1):3–44

Berk RA, Kuchibhotla AK, Tchetgen Tchetgen E. 2021. Improving fairness in criminal justice algorithmic risk assessments using optimal transport and conformal prediction sets. arXiv:2111.09211 [STAT.AP]

Berk RA, Sorenson SB, Barnes G. 2016. Forecasting domestic violence: a machine learning approach to help inform arraignment decisions. *J. Empir. Legal Stud.* 13(1):94–115

Bickel PJ, Hammel EA, O'Connell JW. 1975. Sex bias in graduate admission: data from Berkeley. *Science* 187:394–404

Bielen S, Grajzi P. 2021. Prosecution or persecution? Extraneous events and prosecutorial decisions. *J. Empir. Legal Stud.* 18(4):765–800

Binns R. 2019. On the apparent conflict between individual and group fairness. arXiv:1912:06883v1 [cs.LG]

Bishop CM. 2006. *Pattern Recognition and Machine Learning*. New York: Springer

Blinder A. 2022. Lia Thomas wins an N.C.A.A. swimming title. *New York Times*, March 17. **https://www.nytimes.com/2022/03/17/sports/lia-thomas-swimmer-wins.html#:~:text=Thomas%2C%20a%20fifth%2Dyear%20senior,woman%20to%20win%20an%20N.C.A.A.**

Bognar G, Hirose I. 2014. *The Ethics of Health Care Rationing: An Introduction*. London: Routledge

Bolukbasi T, Chang K-W, Zou J, Saligrama V, Kalai A. 2016. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. arXiv:1607.06520 [cs.CL]

Breiman L. 2001. Statistical modeling: two cultures. *Stat. Sci.* 16(3):199–231

Buja A, Brown L, Berk R, George E, Pitkin E, et al. 2019a. Models as approximations I: consequences illustrated with linear regression. *Stat. Sci.* 34(4):523–44

Buja A, Brown L, Kuchibhotla AK, Berk R, George E, Zhao L. 2019b. Models as approximations II: a model-free theory of parametric regression. *Stat. Sci.* 34(4):545–65

Calmon FP, Wei D, Vinzamuri B, Ramamurthy KN, Varshney KR. 2017. Optimized pre-processing of discrimination prevention. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, ed. I Guyon, U Von Luxburg, S Bengio, H Wallach, R Fergus, et al. N.p.: NeurIPS

Campbell DT, Stanley JC. 1963. *Experimental and Quasi-Experimental Designs for Research*. Chicago: Rand McNally & Co.

Chen Y, Bühlmann P. 2021. Domain adaptation under structural causal models. *J. Mach. Learn. Res.* 22(261):1–80

Chouldechova A. 2017. Fair prediction with disparate impact: a study of bias in recidivism prediction instruments. *Big Data* 5:153–63

Cochran WG. 2007. *Sampling Techniques*. New York: Wiley

Coglianese C. 2021. *Administrative law in the automated state*. Work. Pap. 2273, Legal Sch. Repos., Univ. Pa. **https://scholarship.law.upenn.edu/faculty_scholarship/2273/**

Coglianese C, Lai A. 2022. Algorithm v. algorithm. *Duke Law J.* 71:1281–342

Coglianese C, Lehr D. 2018. Transparency and algorithmic governance. *Adm. Law Rev.* 71:1–56

Cooper AF, Abrams E. 2021. Emergent unfairness in algorithmic fairness–accuracy trade-off research. In *AIES '21: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 46–54. New York: ACM

Corbett-Davies S, Pierson E, Feller A, Goel S, Hug A. 2017. Algorithmic decision making and the cost of fairness. arXix:1701.08230v4 [cs.CY]

Chzhen E, Denis M, Hebiri M, Oneto L, Pontil M. 2020. Fair regression with Wasserstein barycenters. arXiv:2006.07286 [stat.ML]

Diana E, Gill W, Kearns M, Kenthapadi K, Roth A. 2021. Minimax group fairness: algorithms and experiments. arXiv:2011.03108v2 [cs.LG]

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel 2011. Fairness through awareness. arXiv:1104.3913v2 [cs.CC]

Feldman M, Sorelle AF, Moeller J, Scheidegger C, Venkatasubramanian S. 2015. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–68. New York: ACM

Fisher FM, Kadane JB. 1983. Empirically based sentencing guidelines and ethical considerations. In *Research on Sentencing: The Search for Reform, Volume II*, ed. A Blumstein, J Cohen, SE Martin, MH Tonry, pp. 184–93. Washington, DC: Natl. Acad. Press

Freedman D. 2006. *Statistical Models: Theory and Practice*. Cambridge, UK: Cambridge Univ. Press

Friedman JH. 2001 Greedy function approximation: a gradient boosting machine. *Ann. Stat.* 29(5):1189–232

Gillen S, Jung C, Kearns M, Roth A. 2018. Online learning with an unknown fairness metric. In *NIPS'18: Proceedings of the 32nd International Conference on Neural Information Processing Systems*, ed. S Bengio, HM Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, pp. 2600–9. N.p.: NeurIPS

Hardt M, Price E, Srebro N. 2016. Equality of opportunity in supervised learning. arXiv:1601.02413v1 [cs.LG]
Hastie T, Tibshrani R, Friedman J. 2009. *Elements of Statistical Learning*. New York: Springer. 2nd ed.
Hosanna-Tabor Evangelical Lutheran Church and School v. Equal Employment Opportunity Commission, 597 F. 3d 769, reversed (2010)
Imai K, Jiang Z, Greiner DJ, Halen R, Shin S. 2022. *Experimental evaluation of algorithm-assisted human decision-making: an application of pretrial public safety assessment*. Presentation at Royal Statistical Society, Statistics and Law Section, Data Ethics and Governance Section, and Discussion Meetings Committee, virtual meeting, Feb. 8
Imbens GW, Lemieux T. 2008. Regression discontinuity designs: a guide to practice. *J. Econom.* 142(2):615–35
Jung C, Kearns M, Neel S, Roth A, Stapleton L, Wu ZS. 2020. An algorithmic framework for fairness elicitation. arXiv:1905.10660v2 [cs.LG]
Kamiran F, Calders T. 2012. Data pre-processing techniques for classification without discrimination. *Knowl. Inf. Syst.* 33:1–33
Kearns M, Roth A. 2020. *The Ethical Algorithm*. Oxford, UK: Oxford Univ. Press
Kim MP, Ghorbani A, Zou J. 2019. Multiaccuracy: black-box post-processing for fairness classification. In *AIES '19: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–54. New York: ACM
Kleinberg J, Himabindu L, Leskovec J, Ludwig J, Sendhil M. 2018. Human decisions and machine predictions. *Q. J. Econ.* 133(1):237–93
Kleinberg J, Mullainathan SM, Raghavan M. 2017. Inherent tradeoffs in the fair determination of risk scores. In *Proceedings of the 8th Conference on Innovations in Theoretical Computer Science*, ed. CH Papadimitriou, artic. 43. Saarbrücken, Ger.: Schloss Dagstuhl
Kuchibhotla AK, Berk RA. 2020. Nested conformal prediction sets for classification with applications to probation data. arXiv:2104.09358
Kuchibhotla AK, Kolassa JE, Kuffner TA. 2021. Post-selection inference. *Annu. Rev. Stat. Appl.* 9:505–27
Kussner M, Loftus J, Russel C, Silva R. 2018. Counterfactual fairness. arXiv:1703.06856v3 [stat.ML]
Le Gouic T, Loubes J-M, Rigollet P. 2020. Projection to fairness in statistical learning. arXIV:2005.11720v4 [cs.LG]
Lei J, G'Sell M, Rinaldo R, Tibshirani RJ, Wasserman L. 2018. Distribution-free predictive inference for regression. *J. Am. Stat. Assoc.* 113:523
Madaan N, Mehta S, Agrawaal T, Malhotra V, Aggarwal A, et al. 2018. Analyze, detect and remove gender stereotyping from Bollywood movies. *PMLR* 81:92–105
Maity S, Xue S, Yurochkin M, Sun Y. 2021. Statistical inference for individual fairness. arXiv:2103.16714v1 [stat.ML]
Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. 2021. Algorithmic fairness, choices, assumptions and definitions. *Annu. Rev. Stat. Appl.* 8:141–63
Miller JL, Rossi PH, Simpson JE. 1986. Race and gender differences in judgements of appropriate prison sentence. *Law Soc. Rev.* 20(3):313–34
Mohri M, Rostamizadeh A, Talwalkar A. 2018. *Foundations of Machine Learning*. Cambridge, MA: MIT Press. 2nd ed.
Mukherjee D, Yurochkin M, Banerjee M, Sun Y. 2020. Two simple ways to learn individual fairness metrics from data. In *ICML'20: Proceedings of the 37th International Conference on Machine Learning*, pp. 7097–107. N.p.: JMLR
Murphy KP. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press
Nabi R, Shpister I. 2018. Fair inference on outcomes. arXiv:1705.10378v4 [stat.ML]
Nakkiran P, Kaplun G, Bansal Y, Yang T, Barak B, Sutskever I. 2019. Deep double descent: where bigger models and more data hurt. aXiv.1912.02292 [cs.LG].
Pan JP, Yang Q. 2009. A survey of transfer learning. *IEEE Trans. Knowl. Data Eng.* 22(10):1345–59
Pan L, Meng M, Ren Y, Zheng Y, Xu Z. 2021. Self-paced deep regression forests with consideration on ranking fairness. arXiv:2112.06455v1 [cs.CV]
Peyré G, Cuturi M. 2019. *Computational Optimal Transport with Applications to Data Science*. Boston: Now Publ.
Oneto L, Chiappa S. 2020. Fairness in machine learning. arXiv:2012.15816v1 [cs.LG]

O'Reilly S. 2017. Just because you're paranoid… Phillip K Dick's troubled life. *The Irish Times*, Oct. 7. **https://www.irishtimes.com/culture/film/just-because-you-re-paranoid-philip-k-dick-s-troubled-life-1.3243976**

Rawls J. 2001. *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard Univ. Press

Robbins MW, Saunders J, Kilmer B. 2017. A framework for synthetic control methods with high-dimensional, micro-level data: evaluating a neighborhood-specific crime intervention. *J. Am. Stat. Assoc.* 112(517):109–26

Rodolfa KT, Lamba H, Ghani R. 2021. Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy. *Nat. Mach. Intell.* 3:869–904

Romano Y, Barber RF, Sabatti C, Candés E. 2019. With malice toward none: assessing uncertainty via equalized coverage. arXiv:1908:05428v1 [stat.ME]

Romano Y, Patterson E, Candés E. 2019. Conformalized quantile regression. arXiv:1905.03222 [stat.ME]

Rossi PH, Berk RA. 1985. Varieties of normative consensus. *Am. Sociol. Rev.* 50(3):333–47

Rossi PH, Berk RA. 1997. *Just Punishments: Federal Guidelines and Public View Compared*. Piscataway, NJ: Aldine de Gruyter

Rossi PH, Waite E, Bose C, Berk RA. 1974. The seriousness of crimes: normative structure and individual differences. *Am. Sociol. Rev.* 39:224–37

Rothenhäusler D, Meinshausen N, Bühlmann P, Peters J. 2021. Anchor regression: heterogeneous data meet causality. *J. R. Stat. Soc. Ser. B* 83:215–46

Rudin C, Berk U. 2018. Optimized scoring systems: toward trust in machine learning for healthcare and criminal justice. *J. Appl. Anal.* 48(5):449–66

Schökopf B. 2019. Causality for machine learning. arXiv:1911.10500 [cs.LG]

Shannon CE. 1959. Coding theorems for a discrete source with a fidelity criterion. *IRE Int. Conv. Rec.* 7:42–163

Shi C, Wang X, Luo S, Zhu H, Ye J, Song R. 2021. Dynamic causal effects evaluation in A/B testing with a reinforcement learning framework. arXiv:2002.01711v5 [cs.LG]

Singer N, Metz C. 2019. Many facial recognition systems are biased, says U.S. study. *New York Times*, Dec. 19. **https://www.nytimes.com/2019/12/19/technology/facial-recognition-bias.html**

Smith AH, Parish JJ. 2010. *When a Person with Mental Illness Goes to Prison*. New York: Urban Justice Cent.

Starr SB. 2014. Sentencing, by the numbers. *New York Times*, Aug. 10. **https://www.nytimes.com/2014/08/11/opinion/sentencing-by-the-numbers.html**

Stevenson MT, Doleac JL. 2018. *The Roadblock to Reform*. Washington, DC: Am. Const. Soc.

Tibshirani RJ, Barber RF, Candés EJ, Ramdas A. 2020. Conformal prediction under a covariate shift. arXiv:1904.06019v3 [stat.ME]

Tseng G. 2018. Interpreting neural networks. *Towards Data Science Blog*, Nov. 16. **https://towardsdatascience.com/interpretable-neural-networks-45ac8aa91411**

Watson J, Holmes C. 2016. Approximate models and robust decisions. *Stat. Sci.* 11(4):465–89

Wu J, Ma Z, Ramen A, Laudanski K, Hung A. 2021. APOL1 risk variants in individuals of African genetic ancestry drive endothelial cell defects that exacerbate sepsis. *Immunity* 54(11):2632–49

Wu Y, Zhang L, Wu X, Tong H. 2019. PC-fairness: a unified framework for measuring causality-based fairness. arXiv:1910.12586v1 [cs.LG]

Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, et al. 2020. A comprehensive survey of transfer learning. arXiv:1911:02685v3 [cs.LG]