AI and Algorithmic Bias: Source, Detection, Mitigation and Implications

Runshan Fu
Heinz College,
Carnegie Mellon University
runshan@cmu.edu

Yan Huang
Tepper School of Business,
Carnegie Mellon University
yanhuang@cmu.edu

Param Vir Singh
Tepper School of Business,
Carnegie Mellon University
psidhu@cmu.edu

Abstract: Artificial intelligence (AI) and machine learning (ML) algorithms are widely used throughout our economy in making decisions that have far-reaching impacts on employment, education, access to credit, and other areas. Initially considered neutral and fair, ML algorithms have recently been found increasingly biased, creating and perpetuating structural inequalities in society. With the rising concerns about algorithmic bias, a growing body of literature attempts to understand and resolve the issue of algorithmic bias. In this tutorial, we discuss five important aspects of algorithmic bias. We start with its definition and the notions of fairness policy makers, practitioners, and academic researchers have used and proposed. Next, we note the challenges in identifying and detecting algorithmic bias given the observed decision outcome, and we describe methods for bias detection. We then explain the potential sources of algorithmic bias and review several bias-correction methods. Finally, we discuss how agents' strategic behavior may lead to biased societal outcomes, even when the algorithm itself is unbiased. We conclude by discussing open questions and future research directions.

Keywords: artificial intelligence; machine learning; fair machine learning; algorithms; unintended consequence of algorithms; algorithmic bias; algorithmic transparency

1. Introduction

Today, artificial intelligence (AI) and machine learning (ML) are widely used throughout our economy to make important decisions that have far-reaching impacts. For example, employers use ML algorithms to screen job applications, banks and financial institutions use ML tools to assess individual credit-worthiness and assist with loan approval decisions, retailers use recommendation algorithms to recommend items to consumers and/or determine the display order of items in their assortment, doctors use algorithms to guide their medical decision-making, and courts in several states in the United States use an algorithm (COMPAS) for recidivism prediction. Initially, ML algorithms were thought to have the potential to help reduce bias and discrimination that was historically prevalent long before the emergence of "big data," as machines can be designed to be "neutral" and "objective," for example, by not considering sensitive features such as gender and race (i.e., by being gender-blind and color-blind).[1]

However, increasingly, evidence of algorithmic bias has been reported and regularly appears in news headlines. In 2016, ProPublica published an article titled "Machine Bias," which showed that COMPAS, an algorithm used for recidivism prediction, was biased against African Americans (Angwin et al. 2016). Specifically, 44.9% of African American defendants were labeled higher risk but did not re-offend, while this error rate was only 23.5% for white defendants. Meanwhile, 28% of African American defendants were labeled low risk but did re-offend, while this error rate was as high as 47.7% for white defendants. In other words, low-risk African American defendants were almost twice as likely as their white counterparts to be mistakenly labeled high risk, while high-risk white defendants were almost twice as likely as their African American counterparts to be mistakenly labeled low risk. This report raised the issue of algorithmic bias to the public and has attracted considerable attention. Since then, several widely used algorithms have been found to produce potentially biased outputs. For example, David Hansson posted on Twitter that Apple Card, a credit card that Apple created in partnership with Goldman Sachs, declined his wife's request to increase the credit line, but it gave him a credit line 20 times higher, despite the fact that they "filed joint tax returns, live in a community-property state, and have been married for a long time." One Apple representative responded, "I swear we're not discriminating. It's just the algorithm."[2] Another example is the Gender Shades project, which evaluated the performance of commercial gender classification algorithms, including three products developed by Microsoft, IBM, and Face ++.[3] The results showed that all three classifiers had higher error rates for minorities and females. The gap in the error rate

[1] Kleinberg et al. (2018) show that the strategy of blinding the algorithm to race inadvertently detracts from fairness across a wide range of estimation approaches, objective functions, and definitions of fairness.
[2] https://twitter.com/dhh/status/1192540900393705474
[3] http://gendershades.org/

of the IBM classifier between lighter males and darker females was 34.4%, and 93.6% of the Microsoft classifier's misgendered faces were those of darker skin colors (Buolamwini and Gebru, 2018). Similarly, Dastin (2018) found that a hiring algorithm Amazon used in the past favored applicants whose resumes contained words that were more commonly found on men's resumes. Bias in ML seems to be almost ubiquitous, which is concerning for practitioners, policy makers, and academic researchers (Lee, Resnik and Barton 2019).

With these rising concerns, a growing body of literature attempts to understand and resolve the issue of algorithmic bias. The majority of fair ML research considers the technical aspect of algorithmic bias. A community of researchers has considered how to define algorithmic fairness, how to detect/measure algorithmic bias, and how to remove/reduce algorithmic bias in a mathematical/statistical sense. Based on these works, practitioners have developed convenient tools that help detect and mitigate algorithmic bias, such as IBM's AI Fairness 360 toolkit and Microsoft's Fairlearn Python package. These seem to be obvious steps to address the issue of algorithmic bias, but they are complex and challenging. For example, Chouldechova (2017) has shown that different technical fairness definitions usually cannot be satisfied simultaneously. Additionally, in this stream of research, algorithms are often examined in isolation with humans from whom they learn and to whom they are applied, and who ultimately decide whether and how to use the algorithms. In reality, imposing algorithmic fairness (using any fairness notion) will likely alter the incentives of agents who use ML algorithms or who are subject to them. As a result, agents will strategically react to the fairness notion imposed. Without accounting for agents' responses to a fairness requirement, any discussion of its implications is of limited usefulness. Questions about how agents react to fairness requirements, and more broadly, the use of AI and ML, are particularly relevant to policy makers, and answering these questions requires multi-disciplinary perspectives from, for example, economists and social scientists (Manyika 2019).

While the algorithmic bias problem concerns a wide range of contexts, in this tutorial, we use a loan-granting example to illustrate some of the key concepts. Consider a loan-granting decision in which we as decision makers use an ML algorithm to predict whether an applicant will repay the loan, and we only grant loans to those who are predicted to do so. Let $X \in R^M$ be the features of loan applicants, such as income, credit score, homeownership, and the loan amount requested. The sensitive attribute we consider is gender, denoted as $A \in \{0, 1\}$, where $A = 1$ for female applicants and $A = 0$ for male applicants. The decision we need to make is whether or not to grant a loan to each applicant, which we denote as $L \in \{0, 1\}$. An ML algorithm takes data input and produces risk scores, $s = h(X, A) \in [0, 1]$, and the loan-granting decisions are based on the risk scores, i.e., $L = c(s, A)$. For evaluation purposes, assume we also know the actual performance of each applicant if they are granted loans. We denote the performance as $Y \in \{0, 1\}$,

where $Y = 1$ means the applicant will actually repay the loan, and $Y = 0$ means the applicant will eventually default.

In the absence of fairness considerations, our goal is to maximize the prediction accuracy of the algorithm so that we can grant loans to more applicants who will repay and fewer applicants who will default. In the ideal case, we could perfectly predict whether an applicant will default, i.e., $L = Y$ for all the applicants, and we would give loans to every applicant who will repay and reject everyone else. With fairness considerations, we still try to maximize the prediction accuracy but simultaneously attempt to make our decisions "fair."

In what follows, we review the definition of algorithmic bias and the various notions of fairness policy makers, practitioners, and academic researchers have used or proposed. Next, we discuss the challenges in algorithmic bias identification and detection given the observed decision outcome, and we describe methods for bias detection. We then explain the potential sources of algorithmic bias and review several bias correction methods. Finally, we discuss the social and economic implications of ML, emphasizing how agents' strategic behavior may lead to biased societal outcomes even when the algorithm itself is unbiased.

2. Fairness Notions in Machine Learning

In our pursuit of fairness, regulations regarding discriminatory behavior based on protected attributes have been established for several high-stakes domains, such as credit, housing, education, and employment. Current legislation recognizes two doctrines of discrimination: *disparate treatment* and *disparate impact*. Disparate treatment addresses procedural discrimination, and it recognizes liability for formal disparate treatment of similar people from different classes (race, gender, or other sensitive attributes) and intent to discriminate. Meanwhile, disparate impact addresses outcome discrimination, and it recognizes liability for practices with disparate impacts on different classes (Barocas and Selbst 2016).

As ML algorithms are now widely used in decision making in those high-stakes domains, there have been rising concerns about potential algorithmic bias. Before analyzing and resolving the algorithmic bias problem, we need to first define fairness in the context of ML algorithms. Due to the complex nature of the concept, a number of fairness notions have been proposed to accommodate different scenarios and equity goals (Verma and Rubin 2018). In this section, we review several fairness notions in four major categories: unawareness, individual fairness, group fairness, and counterfactual fairness.

Unawareness requires that sensitive attributes are not explicitly used in decision-making processes; therefore, everyone is treated with the same standard regardless of his or her sensitive attribute value. Individual fairness notions reflect the idea that similar people should be treated similarly, and group fairness

notions generally require certain statistics regarding the decisions to be the same across different demographic groups. Counterfactual fairness reflects the idea that a decision regarding an individual should not change if the individual were in a counterfactual world with a different sensitive attribute value.

Unawareness

Also called "blindness" or "anti-classification," unawareness as a fairness notion is consistent with the law that prohibits *disparate treatment*. It simply requires that sensitive attributes are not explicitly used in the decision process. In our example of a loan-granting decision, it means that neither the ML algorithm nor the mapping from scores to decisions can involve the explicit use of the gender, i.e.,

$$h(X, A = 0) = h(X, A = 1) = h(X);\ c(s, A = 0) = c(s, A = 1) = c(s).$$

Unawareness reflects a natural idea of ensuring fairness: remove the sensitive attribute as an input, and the decision maker would not be able to discriminate based on it. With this constraint, everyone is held to the same standard: if two applicants have identical features X, then they would have the same binary outcome on the loans (if c is a deterministic function) or the same probability of obtaining the loans (if c is a probabilistic function).

While unawareness ensures that the sensitive attribute does not explicitly affect the decisions (Corbett-Davies and Goel 2018), algorithms and decision-making processes that satisfy the unawareness requirement often lead to different outcomes across the sensitive attribute groups, thus creating *disparate impact*, as shown in a range of empirical studies (Angwin et al. 2016, Chouldechova 2017, Fuster et al. 2018). This is because sensitive attributes are usually correlated with other features. Thus, even though the sensitive attribute is not directly fed into the ML model, the input features could still carry information about the sensitive attribute. This phenomenon is commonly known as "redundant encoding" (Pedreshi et al. 2008, Dwork et al. 2012). This particular observation suggests that simply prohibiting the explicit use of sensitive attributes may not be sufficient to ensure fairness in ML, and a number of other fairness notions have been proposed to measure the impact (dis)parity of algorithm-based decisions, which we explain in the following sections.

Individual Fairness

Individual fairness notions require the decisions to be fair for any pair of individuals. Dwork et al. (2012) is among the first to popularize the concept, proposing a fairness definition that reflects the idea that "similar individuals should be treated similarly." Formally, let V be the set of individuals and $M(\cdot)$ be the mapping from individuals to probability distributions over outcomes; the decisions are considered fair if the mapping satisfies the (D, d)-Lipschitz property, i.e., for any (i, j) pair in the applicant set,

$$D(M(X^i), M(X^j)) \leq d(X^i, X^j), \forall\, i, j \in V$$

where D is a distance measure for distributions, and d is a distance measure for feature vectors.

In our loan-granting example, this fairness constraint means that if the distance between two applicants (computed under a certain distance measure on X, e.g., Euclidean distance) is d^0, then the distance between their probability of being funded (computed under a certain distance measure on $\Pr(Y = 1)$, e.g., simple difference) should be at most d^0. In other words, under individual fairness notions, if two applicants are similar (in their observed features), their probability of being funded should also be similar.

It is important to note that the similarity of the outcome distributions and that of the individuals are measured by the distance metrics D and d, which are not explicitly defined. Apparently, the validity of this fairness notion crucially depends on the choice of distance metrics. First, as Dwork et al. (2012) have emphasized, the distance metric d should be task specific, i.e., it should measure whether individuals are similar *with respect to a certain task*. In our loan-granting example, it may be justifiable if d measures the similarity of individuals in terms of their default risk or expected repayment performance. However, d might not be appropriate if it measures how similar two individuals are in terms of their productivity or personality. Second, in most cases, the distance metric can only approximate rather than capture the ground truth of the relevant construct (e.g., credit risk), and the quality of the (best) approximation depends on the available features. Thus, an appropriate distance metric requires relevant and comprehensive features. More importantly, the sensitive attributes should not directly or indirectly affect the distance metrics. If individuals in the same sensitive attribute group always have lower distance compared with those in different groups, then the decisions could still be biased (systematically favoring one group over the other), even if the Lipschitz condition is satisfied.

Another implication of the individual fairness notions is that a "cut-off rule" can never be fair. Often, firms use ML algorithms to score individuals and then select those whose scores are above or below a certain threshold for treatment. In our example, we can use an algorithm to estimate applicants' creditworthiness and give loans only to those with high credit scores. Under the individual fairness framework, the difference in the estimated default risk should be considered the "distance" between two applicants. Therefore, individuals whose scores are just above the threshold and those whose scores are just below are similar, and they should be treated similarly. However, with the cut-off rule, one group is always funded, while the other never is. Thus, the individual fairness notions require randomized mappings from features to decisions. Binary decisions, as in our loan-granting example, require that the probability of a positive decision (e.g., granting the loan) should be continuous over the (transformed) feature space. This means that applicants with higher credit scores could have a higher probability of being funded, but in the

final decisions, some applicants with high credit scores may not be funded, while some with lower credit are funded. It is still under debate whether this randomization approach could be truly "fair."

While individual fairness notions are simple and intuitive, the ambiguity and uncertainty that are intrinsic in distance measures, as well as the required randomized mapping that remains controversial, hinder the practical use of such fairness notions, and the way to refine and better formalize individual fairness notions remains an important, ongoing research area.

Group Fairness

Sometimes referred to as the *statistical fairness definition*, *group fairness notions* require equality in certain statistics across different demographic groups. In contrast to individual fairness notions, the group fairness notion does not consider the features of applicants (X) and the mappings from features to decisions. Instead, these notions are defined over certain group statistics that only involve the decision/prediction outcomes (binary decisions or predicted scores) and/or the ground truth label (Y).

Group fairness notions require certain statistics regarding the outcomes to take the same value across demographic groups, and the notions differ in the statistics they focus on. Among the popular group fairness notions, *statistical parity* only considers decision outcomes and requires the same acceptance rate or equal expected score across demographic groups,[4] while other notions also take the ground truth label (L) into consideration. Specifically, *equal opportunity, equalized odds*, and *balance for positive (negative) class* focus on the distribution of outcomes conditional on the ground truth label, while *predictive parity* and *calibration* focus on the distribution of the ground truth label conditional on the outcomes. Moreover, equal opportunity, equalized odds, and predictive parity use the binary decision label (Y) as the outcome measures, while balance for positive (negative) class and calibration use the risk score (s) as the outcome measures. We now move to the motivations and formal definitions of these group fairness notions.

Statistical Parity Also called *demographic parity*, *statistical parity* is one of the most popular fairness notions in the computer science community. It simply requires that decisions (L) and sensitive attributes (A) are independent. In other words, the decisions should have the same distributions across all the demographic groups. In our loan-granting example, it means the approval rate is the same for males and females, i.e.,

$$Pr\,(L\ = 1 \mid A = 0)\ = Pr\,(L = 1 \mid A = 1)$$

[4] Conditional Statistical Parity requires the same acceptance rate, conditional on some legitimate features.

Another version of statistical parity is defined over the algorithm-produced scores s, and it requires equal average scores for different sensitive attribute groups:

$$E(s \mid A = 0) = E(s \mid A = 1)$$

The idea behind statistical parity is that individuals from different demographic groups should have an equal chance of being selected for favorable actions (such as loan granting). This fairness notion implies that, in binary decisions, the demographic composition in the approved set should be identical to the demographic composition in the population.

While it may be desirable to have equal acceptance rates in certain scenarios, such as employment and school admission for diversity or affirmative action, statistical parity is problematic because it ignores an important factor that must be considered in decision making: individuals' qualifications. When one demographic group has more qualified individuals, statistical parity requires us to reject qualified individuals from this group and/or approve unqualified individuals from the other group. It is questionable whether this approach is indeed "fair," and often, it is not aligned with our goals in decision making (Hardt et al. 2016, Dwork et al. 2012).

A variant of statistical parity, *conditional statistical parity*, extends the original definition by allowing decisions to be correlated with certain legitimate features. For example, the debt-to-income ratio may be a strong predictor of loan performance; therefore, making loan-granting decisions based on applicants' debt-to-income ratio should be justifiable and not considered unfair. Formally, conditional statistical parity requires that conditioning on the legitimate feature(s), the distribution of the decisions should be the same across all demographic groups, i.e.,

$$Pr\,(L = 1 \mid Z = z, A = 0) = Pr\,(L = 1 \mid Z = z, A = 1)$$

where Z denotes legitimate features, such as the debt-to-income ratio in our loan-granting decision example. While this fairness notion looks more reasonable in many settings, it leaves significant ambiguity and complexity in the choice of "legitimate features" and may not be able to ensure fairness. A decision maker can choose a feature that correlates with the sensitive attribute and make unfair decisions while satisfying conditional statistical parity. The redlining practice, where governments and businesses decline services to residents in specific neighborhoods, is an example that arguably satisfies this fairness notion but is indeed discriminatory.

Due to the conceptual flaws in statistical parity and its variants, several new group fairness notions have been proposed in recent years. These fairness notions consider not only the final decisions L, but also the true label Y. The underlying idea is that the label Y can serve as the ground truth of qualification. In our example, applicants who will actually repay should be considered qualified applicants and receive loans.

We denote the *predicted* loan performance as $\hat{Y} \in \{0, 1\}$, where $\hat{Y} = 1$ means that the algorithm predicts that the applicant will repay and we will approve the loan, and $\hat{Y} = 0$ indicates otherwise. Note that, in our setting, $\hat{Y} = L$, as we only grant loans to those predicted to repay. Based on the predicted outcome $\hat{Y}$ and the actual outcome Y, we can divide all the applicants into four categories, as shown in Table 1.

Table 1. Categories by Predicted Outcomes and Actual Outcomes

	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	True Positive	False Positive
$\hat{Y} = 0$	False Negative	True Negative

The matrix that reports the number of instances in each category is commonly known as a *confusion matrix*, and many binary prediction performance evaluation metrics are derived from a confusion matrix:

1. True positive rate (sensitivity, recall, hit rate) = true positive / positive = true positive / (true positive + false negative). It measures the fraction of correctly identified positive instances among all positive instances.
2. True negative rate (specificity, selectivity) = true negative / negative = true negative / (true negative + false positive). It measures the fraction of correctly identified negative instances among all negative instances.
3. Precision (positive predictive value) = true positive / (true positive + false positive). It measures the fraction of correctly identified positive instances among all instances identified as positive.

Equal Opportunity and Equalized Odds *Equal opportunity* is one of the most popular fairness notions, and it has received considerable attention since it was proposed in Hardt et al. (2016). The underlying idea of equal opportunity is that *qualified individuals* should have an equal opportunity to obtain favorable outcomes, regardless of their sensitive attributes. In our example, this means that the probability of receiving the loans should be the same for applicants who will actually repay in both male and female groups, i.e.,

$$Pr\,(L = 1|Y = 1, A = 0) = Pr\,(L = 1|Y = 1, A = 1)$$

In other words, *equal opportunity* states that we should correctly identify qualified applicants and approve them at the same rate in both demographic groups. Note that it does not specify any requirement on decisions for applicants who are not qualified. In the same paper, Hardt et al. (2016) also proposed a

stronger fairness notion called *equalized odds*. In addition to the above requirement imposed by *equal opportunity*, *equalized odds* also requires an equal acceptance rate for applicants who will not repay in both demographic groups. Formally, *equalized odds* requires

$$Pr\ (L = 1|Y = y, A = 0)\ = Pr\ (L = 1|Y = y, A = 1), y \in \{0, 1\}$$

In ML terminology, equal opportunity requires an equal true positive rate, and equalized odds requires an equal true positive rate and an equal false positive rate (true negative rate). Thus, while equal opportunity leaves certain flexibility in the prediction performance of the algorithm, equalized odds basically enforces equal prediction accuracy in both demographic groups: if the accuracy is higher for one group, in order to satisfy the equalized odds constraint, we must sacrifice some of the accuracy, such that the true positive rate and the false positive rate in this group match those in the other group.

Equal opportunity and equalized odds address some key drawbacks of statistical parity. First, they take individuals' qualifications into consideration. When one demographic group has both more qualified individuals and more approved individuals, it is no longer ambiguous whether the decisions are biased or justifiable by business necessity under these two fairness notions. Leveraging the knowledge of ground truth labels as qualifications, equal opportunity and equalized odds evaluate decisions for qualified and unqualified individuals separately, in different demographic groups, which is intuitive and logical in most settings. Second, these two fairness notions are often aligned with (unbiased) decision makers' utility goals. Accepting all the qualified individuals and rejecting all the non-qualified individuals (i.e., perfect prediction) maximizes a decision maker's utility but violates statistical parity when proportions of qualified individuals in different demographic groups are not equal. However, equal opportunity and equalized odds are always satisfied under perfect prediction and therefore could potentially motivate better prediction accuracy. Furthermore, Hardt et al. (2016) have argued that equal opportunity and equalized odds "shift the burden of uncertainty in classification from the protected class to the decision maker." Since the two fairness notions require equal accuracy rates, decision makers are motivated to improve the prediction performance for the group with lower prediction accuracy (which is usually the protected group) and therefore improve its welfare.

Balance for Positive Class and Negative Class *Balance for positive class* and *balance for negative class* reflect ideas similar to those of equal opportunity and equalized odds, but instead of measuring distributions of final decisions (L), they focus on the algorithm-produced scores (s). Specifically, balance for positive class requires the average scores of the individuals in the positive class to be the same across all demographic groups. In our loan-granting example, this means that the scores that the ML algorithm produces should have the same expected value for males who will repay and females who will repay, i.e.,

$$E(s \mid Y = 1, A = 0) = E(s \mid Y = 1, A = 1)$$

Symmetrically, balance for negative class requires the average scores of the individuals in the negative class ($Y = 0$) to be the same across all demographic groups. That is, the algorithm-produced scores in our loan-granting example should have the same expected value for males who will not repay and females who will not repay, i.e.,

$$E(s \mid Y = 0, A = 0) = E(s \mid Y = 0, A = 1)$$

The underlying idea is that the algorithm-produced scores should not be systematically higher or lower for one demographic group, conditional on individuals' true label (class), because such a difference cannot be justified after controlling for qualifications. Note that these two fairness notions are defined over the predicted scores, and while the underlying ideas are similar to those behind equal opportunity and equalized odds, there is no clear mapping among these notions. Because of the flexibility in the mapping from predicted scores to final decisions ($c(s, A)$), balance for positive or negative class does not guarantee equal opportunity or equalized odds, and vice versa.

Predictive Parity and Calibration While the previous four definitions focus on the predicted outcomes or scores for individuals with certain actual outcomes, *predictive parity* and *calibration* evaluate the actual outcomes for individuals with certain predicted outcomes or scores. Specifically, predictive parity requires that the proportions of positive instances among the approved individuals are the same across both demographic groups. In our loan-granting example, this means that male applicants who are approved for loans and female applicants who are approved for loans should have the same payback rate, i.e.,

$$Pr\,(Y = 1 \mid L = 1, A = 0) = Pr\,(Y = 1 \mid L = 1, A = 1)$$

In ML terminology, predictive parity is equivalent to equal positive predictive value (or equal precision) across different demographic groups. In a similar spirit, calibration requires that the proportions of positive instances among individuals with the same predicted scores are the same across all demographic groups; in our loan-granting example, this means that the male applicants and female applicants with the same predicted score (s) should have the same payback rate, i.e.,

$$Pr\,(Y = 1 \mid s, A = 0) = Pr\,(Y = 1 \mid s, A = 1), \forall s$$

A stricter version of calibration requires this proportion to be equal to the predicted score, i.e.,

$$Pr\,(Y = 1 \mid s, A = 0) = Pr\,(Y = 1 \mid s, A = 1) = s, \forall s$$

The idea behind predictive parity and calibration is that, for a given predicted outcome or predicted score, the likelihood of an individual being a positive instance should be the same regardless of his or her sensitive

attribute. In other words, the predicted outcomes and the predicted scores should be comparable across demographic groups, and there is no systematic bias with regard to the sensitive attribute in the prediction. The strict version of calibration further requires that "the [predicted] scores mean what they claim to mean, even when considered separately in each [demographic] group" (Kleinberg et al. 2017). In binary classifications, the scores that ML algorithms produce can be interpreted as the predicted probability of being positive instances, and calibration states that the predicted probability should match the actual probability, even in demographic subgroups.

In contrast to the previous four notions, which define fairness as equal treatment for individuals with the same qualifications (actual outcomes), predictive parity and calibration define fairness as similar outcomes for individuals who receive the same treatment. These are two different perspectives. The former focuses on equal access to scarce resources (e.g., credits and loans) conditional on qualifications (actual outcomes), and the latter requires decisions or assigned scores to be justified by actual outcomes. Violations of predictive parity or calibration suggest that, among all the approved applicants or applicants assigned the same scores, those in one demographic group are overall less qualified than those in other demographic groups.

We have reviewed several popular group fairness notions that have been widely discussed and used in the literature. A number of other group fairness notions defined over certain statistics can be derived from a confusion matrix, such as *overall accuracy equality* and *treatment equality* (Berk et al. 2018). There has been a growing body of literature on this family of fairness definitions in recent years, partly because they are simple and intuitive and can be easily verified without access to the algorithms or assumptions on the data. However, group fairness notions cannot ensure fairness on the individual level, as they only put constraints on specific group statistics.

Moreover, it is important to note the inherent conflicts of different group fairness notions. For example, it is straightforward to show that statistical parity and equalized odds cannot be satisfied simultaneously in most cases. By the law of total probability, we have

$$Pr\,(L = 1 \mid A = 0) = Pr\,(L = 1 \mid Y = 1, A = 0) \cdot Pr\,(Y = 1 \mid A = 0) + Pr\,(L = 0 \mid Y = 0, A = 0) \cdot Pr\,(Y = 0 \mid A = 0)$$

$$Pr\,(L = 1 \mid A = 1) = Pr\,(L = 1 \mid Y = 1, A = 1) \cdot Pr\,(Y = 1 \mid A = 1) + Pr\,(L = 0 \mid Y = 0, A = 1) \cdot Pr\,(Y = 0 \mid A = 1)$$

Equalized odds requires that

$$Pr\,(L = 1 \mid Y = 1, A = 0) = Pr\,(L = 1 \mid Y = 1, A = 1) = Pr\,(L = 1 \mid Y = 1)$$

$$Pr\ (L = 1|\ Y = 0, A = 0)\ = Pr\ (L = 1|\ Y = 0, A = 1)\ = Pr\ (L = 1\ |\ Y = 0).$$

When $Pr\ (L = 1|\ Y = 0)\ \neq Pr\ (L = 1|\ Y = 1)$ and $Pr\ (Y = 1\ |\ A = 0)\ \neq Pr\ (Y = 1\ |\ A = 1)$, we have $Pr\ (L\ =\ 1\ |\ A = 0)\ \neq Pr\ (L\ =\ 1\ |\ A = 1)$. This means that, if the acceptance rate is different for qualified and unqualified individuals and the proportions of qualified individuation (base rates) are different across demographic groups, then equalized odds and statistical parity are incompatible.

Such incompatibility is not a single case. Kleinberg et al. (2017) have proved that calibration, balance for positive class and balance for negative class cannot be satisfied simultaneously unless the base rates are the same across all demographic groups or perfect prediction is achieved. Similarly, Chouldechova (2017) has shown that predictive parity and error rate balance (a fairness notion closely related to equalized odds) are incompatible when base rates differ and predictions are imperfect, which holds in all non-trivial applications. Berk et al. (2017) have also shown the incompatibility of the six fairness notions discussed in the paper. These all suggest that it may not be possible to have decisions that are "fair" in all regards (even when focusing on group-level fairness only) under the current social conditions and predictive technology, and there is an inherent trade-off in the choice of fairness notions.

Counterfactual Fairness Counterfactual fairness notions take a different approach to measuring fairness in algorithm-based decisions. The main idea is that a decision for an individual is fair if it remains unchanged in a counterfactual world in which the individual has a different sensitive attribute value. In this counterfactual world, the individual not only has a different sensitive attribute value (A), but may also have different values for other features (X) because the sensitive attribute may causally affect some features, which could further causally affect other features. Therefore, *unawareness*, the strategy of excluding the sensitive attribute from the inputs of decision making, cannot ensure counterfactual fairness, as the changes in other features (as a result of the change in the sensitive attribute) could possibly change the decisions.

To infer the decisions in a counterfactual world, we need to know the values of X when the individuals are in a different demographic group. Thus, the evaluation of counterfactual fairness crucially depends on a set of causal assumptions, which specify how variables affect each other. The causal assumptions are usually represented by a causal graph denoted as $g\ =\ (U, V, F)$, where U is a set of latent variables that we do not observe in the data, V is all the observables and $V \equiv A\ \cup X$, and F is a set of functions specifying how each variable in V is affected by other variables in V and U. With a causal graph g, we can calculate the values of X when the sensitive attribute takes another value, which we denote as X'_g. Counterfactual fairness requires the decisions to be the same in the actual world and in the counterfactual world (Kusner et al. 2017). That is,

$$Pr\ (L\ = 1\ |\ X, A)\ = Pr\ (L\ =\ 1\ |\ X'_g, 1 - A), \forall X, \forall A \in \{0, 1\}$$

In our loan-granting example, consider a female applicant with certain observed features X (such as income, homeownership, and credit history). Under a causal graph g, the applicant would have different observed feature values X_g' if she were a male. Counterfactual fairness requires that the loan-granting decision should be the same for this applicant in the real world, where she is a female with feature values X, and in the counterfactual world, where she was a male with feature values X_g'.

The variants of counterfactual fairness relax this constraint by allowing the sensitive attributes to influence the decisions through certain types of causal paths or by prohibiting only certain types of causal paths that involve the sensitive attribute and the decisions. For instance, Kilbertus et al. (2017) have proposed prohibiting "unresolved discrimination," in which decisions are influenced by the sensitive attribute in a discriminatory manner in the causal graph, as well as "proxy discrimination", in which the decisions are influenced by features that can be used to derive the value of sensitive attributes. Another example is *fair inference,* proposed by Nabi and Shpitser (2018), in which "legitimate" causal paths are allowed. For example, considering an applicant's income level may be justifiable in the loan-granting decisions, even when income level has a causal link to sensitive attributes. Thus, the causal path "A -> income level -> L" is considered legitimate, and the decisions would still be fair if the changes in the decisions are caused by this path only.

Counterfactual fairness notions are individual-level definitions: they evaluate if the decision for an individual is "fair" by comparing the decisions in the actual world and in a counterfactual world. They are distinct from group fairness notions in the sense that the evaluation does not involve decisions for other individuals and does not take prediction accuracy into consideration (Kusner et al. 2017). They are also different from the individual fairness notions we described before, as those individual fairness notions focus on distance measures of individual pairs, while counterfactual fairness notions rely on causal structures. The causal perspective makes counterfactual fairness notions attractive, as they address the fundamental question, "What would the decision be if the applicant were in another demographic group?" However, this particular feature also makes counterfactual fairness notions hard to implement: it is usually difficult if not impossible to verify a causal graph, which often calls the validity of a counterfactual fairness notion into question.

Other Fairness Notions

The fairness notions we have discussed so far mostly focus on binary classifications. While many initial motivating examples, such as loan approval, bail decisions, and school admission, fall into this category, fairness in other types of ML tasks and specific settings has drawn increasing attention in recent years. For example, Elzayn et al. (2019) and Babaioff et al. (2020) have studied the fair resource allocation problem; Ilvento et al. (2020) have proposed the fairness desiderata in online advertisement (sponsored search

auctions); Jabbari et al. (2017) have discussed fairness constraints in reinforcement learning; Yao and Huang (2017) have proposed several fairness metrics for collaborative filtering recommender systems; and Caliskan et al. (2017), De-arteaga et al. (2019), and Jacobs et al. (2020) have explored various ways of measuring bias in semantic representation and natural language processing (NLP).

The meaning and definitions of fairness have been a major topic in the fair ML research community, and the fairness notions we present here are only a sample of those that have been proposed. It is important to note that the goal here is not to have a unified fairness notion that can be applied to all settings or to require algorithm-based decisions to satisfy all fairness requirements. As we have seen, each fairness notion takes a particular perspective and represents a specific fairness consideration, and many notions are not compatible. As the trade-off is inherent, we need to think carefully about the fairness concerns that matter most in different decision settings and choose the appropriate fairness notions accordingly.

3. How to Identify Algorithmic Bias

While a number of fairness notions have been clearly defined, the detection of algorithmic bias is not as straightforward as it may seem. A key challenge with identifying discrimination in algorithms is that the exact algorithm used to make a decision is often not available to policy makers or enforcement agents. Firms typically keep algorithms opaque for social, economic, or technical reasons. Further, algorithms have become increasingly complex. In addition, technical reasons related to interpretability limit an investigator's ability to identify systematic discrimination through a direct analysis of algorithms. As a result, researchers typically rely on analyzing the outcomes of algorithms to identify potential discrimination. In this section, we review some common strategies for identifying algorithmic bias.

Four-Fifths or 80% Rule

One of the earliest methods of detecting bias is comparing the selection rate of different groups. If the selection rate for one group is sufficiently lower than others, we can say this group of individuals is being discriminated against. This method corresponds to the concept of statistical parity described in the previous section, and it is widely used in our society. For example, the Equal Employment Opportunity Commission (EEOC) uses a so-called “four-fifths rule” or “80% rule,” which requires the selection rate of a protected group to be no less than 80% of the selection rate of a regular group (Feldman et al. 2015).

Although the observed differences in the selection rates may be due to bias, it is also possible that they stem from differences in group composition (Simoiu et al. 2017). The four-fifths test assumes that the two groups are similar in terms of qualified members. However, it is highly possible that one group may have disproportionately more qualified members. In our loan-granting example, it is possible that more

male applicants than female applicants will repay the loans. In this case, male applicants may have a higher approval rate even in the absence of discrimination. Several statistical tests have been subsequently proposed to separate the difference in group composition and discrimination, which we discuss next.

Regression Analysis

The most widely used statistical test for discrimination is *regression analysis* (Ayres 2010). Regression analysis is performed to determine the likelihood of favorable (or adverse) decisions across groups based on sensitive attributes. In this analysis, one controls for the variables that capture underlying risk (to account for group composition) and uses sensitive attributes to predict the likelihood of selection. If the regression is correctly specified, a significant, non-zero coefficient for the protected attributes is viewed as the presence of discrimination. However, convincingly claiming discrimination with regression analysis is difficult and sometimes impossible. Sensitive attributes, such as race and gender, are immutable; hence, it is extremely difficult to identify their causal effect on decisions (Greiner and Rubin 2011). However, this type of regression analysis is still widely used to find initial evidence for claiming discrimination in a wide range of scenarios.

Whether a regression analysis correctly identifies discrimination crucially depends on the specifications of the regression. There are two important statistical limitations that can bias the results: *omitted variable bias* and *included variable bias*.

Omitted Variable Bias The *omitted variable bias* is a well-known problem. If a researcher fails to include a variable that *correlates* with a sensitive variable upon which the decision maker actually based her decision, then the regression can erroneously indicate that the decision maker treated a protected group differently than the regular group (Ayres 2010, Jung et al. 2018, Lakkaraju et al. 2017). For instance, in the loan-granting example, a decision maker may use payment history, which correlates with gender, in loan-granting decisions. If the investigator does not include payment history in his regression, he may incorrectly conclude that the decision maker discriminates based on gender. It is almost impossible to include all variables that may affect the risk scores, and several of these omitted variables may be weakly correlated with sensitive attributes, which may bias the results and lead to incorrect conclusions.

Included Variable Bias Another statistical issue that can bias the statistical inference is the *included variable bias*, which occurs when the investigator inappropriately includes non-justified variables that may be correlated with protected attributes (Ayres 2005, Ayres 2010, Jung et al. 2018). The following is an extreme example: in the case of hiring, the investigator may include the presence of facial hair as one of the variables in the regression to predict the likelihood of hiring. This variable is highly correlated with

gender and in itself should have no impact on hiring decisions. Including it in the regression would incorrectly reduce the effect of gender in the regression.

Outcome Test

Often, we cannot observe all the factors the decision maker has taken into consideration. Thus, "omitted variable bias" and "included variable bias" are beyond mere theoretical threats. Researchers have proposed a novel bias detection method: the *outcome test,* which is based not on the rate at which decisions are made (e.g., the loan approval rates), but on the success rate of those decisions (e.g., the payback rate of the approved applicants; Becker 1993). For instance, in the loan-granting example, even if minorities are less creditworthy than whites, the minorities who are granted loans, absent discrimination, should still be found to repay their loans at the same rate as whites who are granted loans. This method corresponds to the idea of predictive parity described in the previous section. If minorities have a higher repayment rate than whites, this suggests that the decision makers apply a double standard, granting loans only to exceptionally qualified minorities (Simoiu et al. 2017). The outcome test's strength lies in the fact that it does not require any knowledge of the factors the decision maker has taken into consideration nor of the algorithm itself. As a result, it is immune to the omitted variable bias and the included variable bias. However, outcome tests are known to suffer from the *infra-marginality problem* (Ayres 2010, Simoiu et al. 2017).

The infra-marginality problem is that, even absent discrimination, the repayment rates of minority and white loan recipients may differ because of differing underlying (true) risk distributions (Simoiu et al. 2017). For example, assume the risk distribution of whites and that of minorities are two-point distributions. A fraction of the white population has a repayment probability of 70%, and the other fraction has a repayment probability of 93%. Meanwhile, for the minorities, a fraction of the population has a repayment probability of 50%, and the remaining has a repayment probability of 96%. If the decision maker grants loans to anyone whose repayment probability is greater than 90%, then the outcome test would incorrectly conclude that the decision maker requires higher standards from minorities to grant loans.

The infra-marginality problem occurs because, in the outcome tests, we compare the average performance of all individuals from the two groups who are beyond the threshold. However, to check if a decision maker applies the same standard to different groups, we should compare the performance of only individuals who are at the margin (i.e., who just pass the threshold). Recently, Simoiu et al. (2017) have proposed a threshold test that mitigates the problem of infra-marginality by jointly estimating decision thresholds and risk distributions.

Benchmarking

Establishing algorithmic discrimination is essentially an econometric identification problem. It is particularly challenging because the variables of interest — race, gender, and other protected attributes — are time invariant. Thus, econometric approaches that establish identification by exploiting overtime variation are not applicable. Nonetheless, other related methods can help identify algorithmic discrimination. A statistical approach that can be applied to detect algorithmic discrimination but is rarely used is *benchmarking*.

The key methodological challenge of benchmarking analysis is identifying the race/gender distribution of the at-risk or benchmark population. There are a number of creative ways researchers in criminology have established benchmark populations in the case of police stops and searches. The goal of these studies is to identify whether the race of drivers plays a role in whom police stop or search. To identify the benchmark population, researchers need to find the at-risk population in situations where race was unlikely to play a role in their identification of being at-risk. McConnell and Scheidegger (2001) have used the stops initiated by aerial patrols, and Lange et al. (2011) have used traffic tickets/stops based on radars and cameras. Race is unlikely to play a role in initiating these stops or tickets. As a result, the race distribution of these individuals should act as a useful benchmark.

Establishing the right benchmark population is challenging. For example, in the example related to police stops, the identified benchmark population may not be the right benchmark if the race composition of commuters in places where the radars/aerial patrols/cameras are deployed is different from places where police are physically present.

Identifying algorithmic bias is a challenging problem. While researchers have made progress on this issue in recent decades, a fool-proof and cost-efficient method is still missing. Experiments are one potentially costly way in which algorithmic discrimination can be more cleanly studied. For example, in the loan-granting case, an experiment would be to randomly grant loans to a number of individuals to learn the true distribution of at-risk populations. However, such experiments would be quite costly and may not always be feasible.

4. Source of Algorithmic Bias

Before addressing the question of how to address algorithmic bias, it is important to understand why ML algorithms may exhibit bias. The general view is that an algorithm itself is usually not biased in any meaningful way (unless it is coded to be biased), but it may pick up and amplify potential biases in the input data. Machine learning algorithms are designed to find statistical corrections in the data; thus, if the input data carries social biases, the trained algorithm is likely to reflect the same biases (Nkonde 2019). As

Hardt (2014) puts it: "data [is] a social mirror." In this section, we discuss how input data could be biased and why it may cause algorithmic bias.

Biased Labels

Perhaps the most obvious potential bias is in the predictive labels, which are used as the ground truth for the predictive objectives in training ML algorithms. If they are biased, then, not surprisingly, the trained algorithms will try to mimic such biases.

The biases in labels can come from a variety of sources. First, the biases may directly originate in human decision makers when the labels are historical human decisions. In the loan-granting example, if the ML algorithm is trained to predict who should be approved based on historical approval decisions, rather than historical loan outcomes, then the machine prediction may be encoded with biases in past decisions. Barocas and Selbst (2016) provide a real-life example of biased results from biased labels: a hospital in the UK once built a computer program to help automate medical school admission, and the program was developed based on the previous admission decisions. However, those previous decisions were biased against females and racial minorities because of human biases. The automated decisions the program produced were thus also biased against these demographic groups. Barocas and Selbst (2016) cite the editors at the *British Medical Journal*: "(T)he program was not introducing new bias but merely reflecting that already in the system." They conclude that such automations "would turn the conscious prejudice or implicit bias of individuals involved in previous decision making into a formalized rule that would systematically alter the prospects of all future applicants."

Second, the biases may be introduced in the manual labeling process. In many ML tasks, labels are not available, and we may hire people (e.g., from Amazon Mechanical Turk) to manually label instances. In this case, the level of bias encoded in the labels depends on individual labelers, which may be more difficult for the decision makers to control. For instance, in the loan-granting example, if we do not observe the actual payback status and ask people to label the creditworthiness of each applicant, then females may be systematically labeled less creditworthy if the labelers are biased against women. In practice, decision makers may aggregate the labels multiple labelers create in the hope that individual biases can cancel each other out and to some extent reduce the biases in the final labels.

Finally, the biases in labels may arise in the choice of proxies for the ground truth. Obermeyer et al. (2019) have discussed an example of such biased labels. They found significant racial bias in a commercial prediction algorithm that is widely used in the healthcare industry: at the same level of predicted risk, black patients are considerably sicker than white patients in reality. They explored the mechanism of the bias and realized that the predictive objective of the algorithm is healthcare costs instead of illness.

While the cost seems to be an effective proxy for illness severity, it does not account for the unequal access to healthcare resources; rather than being less sick, some patients are low cost because they cannot afford or have limited access to medical treatment. In this case, the algorithm is provided biased objectives, and it is unsurprising that its outputs are biased in a similar way.

Similar cases can be found in many other settings where algorithms are trained with biased labels. Thus, it is crucial to evaluate if the data used as the objective of the prediction (i.e., the label) encodes any explicit or implicit social bias when training algorithms.

Imbalanced Representation

Even in the ideal case where labels are completely free from bias, algorithms could still produce biased predictions. Another important cause of algorithmic bias is the imbalanced representation of different demographic groups.

The protected groups (e.g., racial minorities, females) are usually underrepresented in data for various reasons. First, some protected groups, such as racial or ethnic minorities, naturally have lower populations compared to regular groups. Second, in many cases, training data are samples selected by previous decisions. For example, in making the loan-granting decision, we train an ML algorithm that predicts loan performance based on historical loan performance. Note that we do not observe performance for the loans that have not been approved. Thus, our training dataset consists of approved loans only. If females have been discriminated against in the previous loan approval decisions, then they would be underrepresented in the data. Moreover, there may be self-selection. Even if the previous decisions are completely fair, people who never apply for loans could not be approved, and therefore would not be included in the data used to train algorithms. The reality may be that individuals in the protected groups (e.g., females) have few resources to support a loan application. Crawford (2013) has discussed an interesting example of self-selection: the city of Boston once released a mobile app that used accelerometer and GPS data from personal smartphones to detect potholes and report them to the city. Innovative as the approach is, the problem is that not everyone downloaded the app and participated in the project. People with low income are less likely to have smartphones and were therefore unable to opt in to provide data, and they may live in neighborhoods with more serious pothole problems. In this case, a significant number of people who needed the resources most may be missing in the data.

How does imbalanced representation bias algorithm predictions? We know that, in general, with all else being equal, more data tends to give us more accurate predictions. Imbalanced representation means we have fewer observations for the protected groups compared to other groups, which suggests that the predictions for the protected group may be less accurate. This would not happen if individuals in all the

demographic groups were from one homogeneous population, in which case each group could benefit from observations in other groups. However, in reality, cultural differences and social inequity mean that different demographic groups usually have different statistical patterns in the observed features and the prediction objectives, and what algorithms learn from one group may not be applicable to another. When a group-blind algorithm is trained to minimize overall errors, it will fit the majority group when it is not flexible enough to fit all the groups (Hardt 2014, Chouldechova and Roth 2018). From the previous section on the definitions of fairness, we know that prediction error disparity is a major consideration in algorithmic bias. Therefore, less accurate predictions for the protected groups usually mean unfair predictions.

The overrepresentation of protected groups could also be a problem. For example, police may patrol certain areas more often than other areas, and as a result catch more crimes in the former. As a result, in historical crime data, these areas are overrepresented. It is not hard to imagine that, if we build an algorithm to predict the crime rate, these areas would have significantly higher crime risk, even if the actual crime rate is equal to or lower than other areas.

Data Quality Disparity

A data quality disparity is also a potential cause of algorithmic bias. The meaning of this disparity is twofold. First, the input data may be less complete, accurate, or timely in protected demographic groups due to efficiency gaps in the data collection process and data maintenance (Executive Office of the President 2016; Barocas and Selbst 2016). This means there could be more noise in input features for people in protected groups, which makes prediction more difficult for these people. Second, the predictiveness of the same feature may differ across demographic groups; thus, as in predictive power, the data quality could be lower in certain groups, even though same features are used for everyone. For example, education is usually an important factor used to distinguish competent job applicants from incompetent ones. While the experience of attending a reputable university may help us draw distinctions among people in certain demographic groups, it may be less helpful in the protected groups, as most of the people in these groups have no access to reputable universities regardless of their competency level. Features that better account for pertinent statistical variation among members of protected class are usually more expensive to collect and are therefore ignored (Barocas and Selbst 2016). Both types of data quality disparity could lead to less accurate predictions for protected groups and hence biased outcomes.

5. Correction Methods

As several studies have demonstrated the common presence of algorithmic bias and its causes, a natural question is whether we can reduce or even remove potential bias in algorithms. It is clear that most algorithms cannot be completely bias-free because the different perspectives of fairness captured in various

fairness notions cannot all be satisfied simultaneously in any non-trivial case, as we discussed in Section 2. Nonetheless, a number of methods have been proposed to mitigate algorithmic bias problems. These correction methods generally fall into one of the three categories: pre-processing (data), adding fairness constraints, and post-processing (predictions).

Data Pre-Processing

This approach to corrections aims to remove or mitigate potential bias by transforming data such that it contains no information of sensitive attributes while retaining the maximum possible task-relevant information (Chouldechova and Roth 2018). Recall that even when sensitive attributes are not included in the input features, algorithms could still infer sensitive attributes from other features due to redundant encoding. Therefore, the transformed fair representations of the original need to be statistically independent of the sensitive attributes (Zemel et al. 2013). One way to achieve this is to preserve only rank information and remove dependence with the sensitive attributes in each feature (Feldman et al. 2015). This approach only works for continuous variables; Johndrow and Lum (2017) and Fu et al. (2020) have extended this framework by proposing additional transformation methods that can handle other types of variables (e.g., discrete variables, binned variables). For example, if income and homeownership are strong predictors of loan payback status, and they are highly correlated with gender, then we can map the two features to a space orthogonal to gender before feeding them into the ML algorithm, which prohibits the algorithm from inferring gender information from the two features.

Another stream of methods that removes statistical dependence between input features and sensitive attributes is adversarial learning. The basic framework is to simultaneously learn a predictor and an adversary in a neural network. The predictor tries to predict the labels, while the adversary tries to model the sensitive attributes. The goal is to maximize the predictor's ability to correctly predict the labels while minimizing the adversary's ability to predict the sensitive attributes. In this way, we can obtain representations of the original data that maintain relatively high predictive power while reducing potential biases with respect to the sensitive attributes.

Adding Fairness Constraints

Another approach of correction is to add fairness constraints while training ML algorithms. This approach is less general than data pre-processing, as it requires specification that may differ for different ML models, while processed data that is independent of sensitive attributes can essentially be used in any algorithm without the risk of statistical disparity.

For example, Bechavod and Ligett (2017) have proposed a method to mitigate bias in binary classification, where they aim to achieve a similar false positive rate as well as a similar false negative rate in two demographic groups. The idea is simple: they add a penalty term that measures the disparity of the two rates into the objective function of the ML algorithm. As a result, while the algorithm searches for parameters that minimize prediction errors, it also balances the disparity in the false positive rate and the false negative rate. Similarly, Kamiran et al. (2010) have proposed a method to remove bias in a decision tree classifier by adding fairness constraints that change the splitting criterion and the pruning strategy.

Post-Processing Predictions

The last strategy is to post-process prediction results. Doing so is particularly useful in settings where input data and ML algorithms are not available. In these cases, we can still try to correct algorithmic bias by processing final prediction results.

Most of the post-processing techniques can be boiled down to a group-specific threshold setting. For example, suppose we want to achieve equal opportunity. Due to sampling bias and data quality disparity, our predictions of loan performance are worse for females; thus, if we apply the same threshold to both males and females and only approve loans for applicants whose predicted probability of repayment is higher than the threshold, we would end up approving a lower percentage of females who would actually repay and violate the equal opportunity constraint. If no information other than the predicted scores is provided, we can lower the threshold for females such that the true positive rate of the female group equals that of the male group.

While this approach is simple and intuitive, a meaningful solution does not always exist. For instance, if we aim to achieve equalized odds, we may not have a feasible solution when prediction accuracy differs too much for different demographic groups, or the only feasible solution may not be sensible, e.g., approving only a very small number of applicants. In addition, some fairness constraints, such as individual fairness notions, are difficult to achieve with post-processing techniques.

6. Social and Economic Implications of Machine Learning

The papers we have reviewed in the previous sections focus on the technical aspect of fairness issues in algorithmic bias. For more comprehensive reviews of these topics, we refer interested readers to Barocas et al. (2019). In this section, we examine the social and economic consequences of the use of ML algorithms through the economic lens.

Human vs. Algorithm

In most technical work on algorithmic bias, the focus has been on the absolute bias in the machine outputs. Economists are often interested in comparing bias between multiple approaches, or counterfactuals (Cowgill and Tucker 2017), and determining whether algorithms can improve the *status quo* because, if a decision is not made by the machine, it will most likely be made by a human. Therefore, it makes sense to benchmark algorithms against human judgements. However, it is not as easy to achieve a fair comparison between human and algorithm as it may appear. One significant challenge is the "selective labels" problem in the data, which means that the process generates only a partial labeling of the instances, and the decision maker's choices determine which instances have labels at all (Lakkaraju et al. 2017). Lakkaraju et al. (2017) have proposed a "contraction algorithm" to compare human and machine decisions in the presence of the selective labels problem. Kleinberg et al. (2018) have applied this comparison algorithm to bail decisions and shown that machines can bring potentially large welfare gains — crime reductions up to 24.7% with no change in jailing rates, or jailing rate reductions up to 41.9% with no increase in crime rates — and these gains can be achieved while simultaneously reducing racial disparities. Bartlett et al. (2019) examine bias in consumer lending decisions by individual human examiners using administrative data from a high-cost lender, and they found significant bias against both immigrant and older loan applicants when using the firm's preferred measure of long-run profits, but no evidence of bias when using a short-run measure used to evaluate examiner performance. The authors explain that the bias in their setting was caused by the misalignment of firm and examiner incentives. They follow the approach in Kleinberg et al. (2018) and show that a decision rule based on ML predictions of long-run profits could simultaneously increase profits and reduce bias. Fu et al. (2020) have conducted a similar comparison between the machine and the crowd, instead of individual investors, in the P2P lending context, and shown that a reasonably sophisticated algorithm can more accurately predict the default risk of loan listings than the crowd, and both the borrowers and lenders can simultaneously benefit from the machine. The authors also propose a general debias method and show that the debiased algorithm can increase welfare while minimizing the bias.

Cowgill (2018) has performed a field experiment to compare an ML algorithm and humans in resume screening and granting interview opportunity for white-collar jobs. The author found that the machine-chosen candidates outperformed those selected by human screeners on nearly all measures, such as passing an interview and eventually receiving the job offer, as well as their productivity once hired as employees. However, human benchmarking does not imply that, as long as machines can improve upon human decisions in reducing bias, we should be satisfied and there is no need to undertake actions to further reduce bias in decisions (Cowgill and Tucker 2020).

Compared with human bias, algorithmic bias seems more visible; machines do not lie or hide their biases. This is perhaps generally positive because any bias will be discovered earlier, firms can be held

accountable for it, and improvements can be developed sooner to reduce and eliminate bias. However, the higher visibility of algorithmic bias can also lead to algorithmic bias being disproportionately reported and punished, thus reducing firms' incentive to adopt algorithms, even in cases where algorithms can reduce bias. This is evident in a survey of business decision makers conducted by Cowgill et al. (2020). The authors found that managers are more likely to abandon AI for manual review by humans after they see algorithmic fairness critiques due to concerns about lawsuits and negative PR. The consequence is that firms then return to the "manual" mode, and the benefit and improvement algorithms can bring in reducing bias cannot be realized. In relation to human benchmarking, it is useful to perform such benchmarking to show how algorithms compare with humans regarding the bias in their decisions, and to "credit" algorithms for the improvement they produce. We do not suggest that "doing better than humans" is sufficient. Regulating algorithmic bias and raising awareness are still important. The key is to balance the incentive to adopt algorithms and the incentive to reduce and eventually eliminate bias in algorithms and algorithm-assisted decisions.

Incentives and Economic Mechanisms

In most technical research on algorithmic bias, the data used to train and test the algorithm is treated as given; when proposing new fairness notions and policy changes, the behavior of agents who use the algorithm or are subject to algorithms is often assumed to remain unchanged. However, in reality, agents are likely to strategically react to changes in the algorithms and/or restrictions on how algorithms are used. Assuming away agents' strategic reactions may lead us to an incorrect and misleading conclusion. Cowgill et al. (2020) have provided an extreme example: a large negative penalty on detected algorithmic bias would lead to organizations completely abandoning algorithms and reverting to the manual process, in which bias is less visible and difficult to detect.

Fu et al. (2019) have developed a theoretical model that endogenizes firms' investments in learning: to learn more accurately about the outcome of interest, firms need to incur costs to collect high-quality data, build the relevant infrastructure, develop and update ML models, etc. The incentive for firms to invest in learning is that more accurate predictions can allow them to make better decisions that increase their profits. Thus, a rational firm will choose the optimal amount of learning effort that maximizes its net profits. The authors show that some new notions of machine fairness, such as equal opportunity, that are proposed as better alternatives to equal treatment (unawareness), the requirement by law, while conceptually appealing, can make everyone worse off, including the very group they aim to protect. This is because enforcing fairness in the output of the algorithm and the decisions made based on it essentially imposes constraints on an optimization problem, and such constraints can negatively affect the benefits from the algorithms.

Compared to the current law, the requirement of "equal opportunity" further reduces the benefits from a more accurate algorithm for a firm. As a result, profit-maximizing firms reduce the investment in learning when greater accuracy is otherwise desired and may grant fewer loans to both demographic groups. This paper highlights the importance of considering stakeholders' strategic behavior when developing fair ML algorithms and indicates that, after taking stakeholders' strategic reactions into consideration, some of the "fairness" requirements, if turned into law, may have unintended consequences.

Moreover, unbiased algorithms may sometimes lead to unintended societal outcomes. An example is Lambrecht and Tucker (2016), who conducted a field test to examine how an ad delivery algorithm used by a popular advertising platform delivers an ad promoting job opportunities in the science, technology, engineering, and math (STEM) fields in 191 countries; the ad is explicitly intended to be gender neutral in its delivery. The authors found that the STEM ads are shown more often to men than to women, and the difference is largest for individuals in their prime career years. They examined different potential explanations for this difference and found that it was not driven by the algorithm learning from biased data but by the economics of ad delivery. Women are less likely to see these ads for STEM jobs because competition from other advertisers for young females' attention is more intense; ad auctions for female "eyeballs" contain more bidders and thus have higher clearing prices. Therefore, an algorithm that optimizes cost-effectiveness will deliver ads that were intended to be gender neutral in an apparently discriminatory manner because of the crowding-out effect. These studies emphasize that economic forces may lead to apparently discriminatory outcomes even if the algorithm is designed to be fair, and they highlight the importance of considering agents' strategic behavior and the potential tension between reducing bias and using economic mechanisms to efficiently allocate resources through algorithms (Cowgill and Tucker 2020). In addition, it is useful to note that the trade-off between fairness and efficiency is not unique to algorithms; it existed even before the emergence of algorithms, and researchers have proposed methods that account for fairness and efficiency in optimization (e.g., Bertsimas et al. 2012 and Hooker and Williams 2012). However, the use of algorithms may have made this trade-off more explicit.

Fu et al. (2020) have demonstrated another mechanism whereby an unbiased algorithm has unequal effects on different groups of individuals. The authors studied the impact of "Zestimate," a machine-generated estimate of property value provided by Zillow, the leading real estate and rental website in the United States.[5] The authors examined data on Zestimate and housing sales and show that Zestimate, although not biased, may promote inequality. Zestimate relies on a large number of features on a property to be accurate. The authors found that, properties in richer neighborhoods report more features than properties in poorer neighborhoods, which often have a high concentration of African American households.

[5] https://www.statista.com/statistics/381468/most-popular-real-estate-websites-by-monthly-visits-usa/

As a result, Zestimate is fairly accurate in richer neighborhoods, while it is less accurate in the poorer neighborhoods. Properties that are seriously undervalued by Zestimate are less likely to be sold, and in anticipation of this, owners of those properties are less likely to list these properties on the market in the first place. In contrast, properties that Zestimate overvalues are more likely to enter the market. Overall, Zestimate promotes inequality by pushing more properties in poorer neighborhoods out of the market. This is an example of the unequal impact of algorithm-generated predictions that reflect larger, underlying social problems than algorithmic bias as defined above.

Some researchers have suggested that the use of algorithms is also likely to facilitate collusion, especially in the pricing context. Ezrachi and Stucke (2016) have argued that pricing algorithms can be "a recipe for tacit collusion" that can hurt consumers' welfare. Calvano et al. (2018) have constructed AI pricing agents and let them repeatedly interact in controlled environments where the agents play a repeated pricing game with simultaneous moves and full price flexibility. They found that even relatively simple pricing algorithms can systematically learn to play sophisticated collusive strategies. Brown and MacKay (2019) have shown that collusion is not necessary for algorithms to generate higher prices, and competition in pricing algorithms can also lead to higher prices. On the other hand, Miklós-Thal and Tucker (2019) have developed a theoretical model and shown that better demand prediction can lead to lower prices and higher consumer welfare because of competing firms' ability to better tailor prices to demand conditions and their increased incentive to deviate to a lower price in time periods of more accurate forecasts of high demand.

Transparency

Algorithmic transparency has been in increasing demand to improve accountability and trust in algorithms and to combat algorithmic bias and discrimination. With transparent algorithms, bias detection, as discussed in Section 3, becomes much easier. For example, the European Union's General Data Protection Regulation (GDPR) dictates that, whenever personal data is subject to automated decision making, people have "the right to obtain human intervention on the part of the controller," or the right to explanation.[6] However, one potential drawback is that transparency can make it even easier for firms to collude because the observability of competitors' past, current, and future actions — in this case, through the algorithms they use — can help firms coordinate and collude (Bourveau et al., 2019). In addition, potential gaming behavior as well as privacy and security concerns are also commonly cited arguments against algorithmic

[6] See "Algorithmic transparency and the right to explanation: Transparency is only the first step" https://www.apc.org/en/blog/algorithmic-transparency-and-right-explanation-transparency-only-first-step

transparency (Ford and Price 2016, Wachter et al. 2017, and Kroll et al. 2017). These studies highlight the potential costs of requiring algorithmic transparency.

However, Wang et al. (2020) have built a game-theoretic model to explain how strategic individuals behave under a firm's transparent and opaque algorithms respectively in the hiring context, and they have shown that firms are better off making algorithms transparent than keeping them opaque. The authors consider a generic algorithm a firm uses to evaluate job applicants. The input of the algorithm contains a set of causal features, which are costly to improve and have a causal impact on job applicants' ability, and a set of correlational features, which are costless to improve and only correlated with unobservable causal features and therefore are predictive of job applicants' ability. However, manipulating correlational features does not affect a job applicant's ability. They show that making algorithms transparent may improve their predictive performance in the presence of strategic users. On the one hand, when the algorithm is made transparent, low-type individuals will "game" the system by improving their correlational features, and as a result, both high- and low-type individuals will have the same value in those features, and the predictive power of the correlational features decreases or even disappears. This negatively affects the firm's payoff. On the other hand, anticipating that they will lose their advantage on the correlational features, high-type individuals have a stronger incentive to improve on their causal features to separate them from the low-type individuals, increasing the predictive power of the causal features and their ability. This positively affects the firm's payoff. When the latter effect outweighs the former, the transparent algorithm is more desirable.

Regulations

With the deep roots and complex nature of the algorithmic bias problem, there have been calls for regulations that reconcile different fairness notions and establish practice standards. Several organizations and companies have published guidelines for implementing trustworthy and ethical AI, with fairness as an essential component. Notable examples include the European Ethics Guidelines for Trustworthy AI, the Organization for Economic Co-operation and Development (OECD) Principles on AI, the White House's Guidance for Regulation of Artificial Intelligence Applications, Google's Perspectives on Issues in AI Governance, and IBM's Everyday Ethics for Artificial Intelligence. For a more comprehensive list of guidelines on ethical AI, we refer interested readers to Jobin et al. (2019). These guidelines recognize the importance of identifying and mitigating algorithmic bias, and some provide practical advice on how to design fairness-aware algorithms.

There have also been attempts to formalize anti-algorithmic-bias practice into law and address the gap in the current legislation (Thryft 2019, Kim 2016). In December 2017, the New York City Council passed "a local law in relation to automated decision systems used by agencies"; it is the first law on

algorithm accountability and fairness in the United Sates. It requires the creation of a task force[7] to monitor agency-automated decision systems and provide recommendations. In April 2019, a federal bill called The Algorithmic Accountability Act (S. 1108, H.R. 2231) was introduced. If passed, it will require certain companies to evaluate "highly sensitive" automated systems on "accuracy, fairness, bias, discrimination, privacy, and security" and address the issues identified in a timely manner. The companies covered by the bill include those that 1) have greater than $50 million in average annual gross receipts, 2) possess or control personal information on more than 1 million consumers or devices, and 3) act as data brokers that collect and trade personal data. Similar state-level bills include the New Jersey Algorithmic Accountability Act (NJ A5430) and Washington State bills S.B. 5527 and H.B. 1655.

7. Concluding Remarks

In this tutorial, we reviewed the current literature on algorithmic bias. We highlighted the fact that, due to the complex nature of the concept, a number of fairness notions have been proposed to accommodate different scenarios and equity goals; each fairness notion takes a particular perspective and represents a specific fairness consideration, and some notions are incompatible with each other (i.e., cannot be satisfied simultaneously). We noted this inherent trade-off and stressed the importance of determining the most significant fairness concerns on a case-by-case basis. There may be an opportunity to use an economics framework that explicitly models the trade-off to guide the choice of fairness notions. We reviewed existing methods for bias detection, noted their limitations, and highlighted the challenge in detecting bias in opaque algorithms without knowing either how the algorithm works or the ground truth in the population. Experiments may offer the opportunity to cleanly detect algorithmic bias through benchmarking, but it is costly. A potential future research direction may be to study the optimal design of such experiments to identify algorithmic bias while minimizing cost. Finally, we reviewed the literature that examines algorithmic bias through an economics lens and that investigates the social and policy implications of algorithmic bias, or more broadly, the use of algorithms. We argued that it is important to account for agents' reactions to any imposed fairness constraints when studying the policy implications of that fairness notion. Economic modeling would be useful to study agents' strategic behavior and gain a holistic view of the effect of imposing the fairness notion. In addition, there is ample literature in economics, psychology, and sociology on the topic of human bias and discrimination, and many of the underlying issues in algorithmic bias are familiar to those in human bias. Since one important source of algorithmic bias is human bias that is encoded in the data used to train the algorithm, it may be interesting to explicitly model how humans are biased, how the human bias is reflected in the data, and how the machine inherits or even amplifies the

[7] https://www1.nyc.gov/site/adstaskforce/index.page

human bias. An understanding of these processes may be able to guide the effective development and choice of the bias correction method.

References

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

Ayres, I. (2005). Three tests for measuring unjustified disparate impacts in organ transplantation: The problem of "included variable" bias. *Perspectives in Biology and Medicine*, *48*(1), S68-87.

Ayres, I. (2010). Testing for discrimination and the problem of "included variable bias", mimeo, Yale Law School.

Babaioff, M., Nisan, N., & Talgam-Cohen, I. (2019). Fair allocation through competitive equilibrium from generic incomes. In *FAT* (p. 180).

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104, 671.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning. fairmlbook.org.

Bechavod, Y., & Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Becker, G. S. (1993). Nobel lecture: The economic way of looking at behavior. *Journal of Political Economy*, *101*(3), 385-409.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, https://doi.org/10.1177/0049124118782533.

Bertsimas, D., Farias, V. F., & Trichakis, N. (2012). On the efficiency-fairness trade-off. *Management Science*, *58*(12), 2234-2250.

Bourveau, T., She, G., & Zaldokas, A. (2019). Corporate disclosure as a tacit coordination mechanism: Evidence from cartel enforcement regulations. *Available at SSRN 2954382*.

Brown, Z. and MacKay, A. (2020), Competition in pricing algorithms. *Available at SSRN 3485024.*

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334), 183-186.

Calvano, E., Calzolari, G., Denicolò, V., & Pastorello, S. (2018). Artificial intelligence, algorithmic pricing and collusion. *Available at SSRN 3304991*.

Chen, I., Johansson, F. D., & Sontag, D. (2018). Why is my classifier discriminatory? In *Proceedings of 32nd International Conference on Neural Information Processing Systems* (pp. 3543-3554).

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2), 153-163.

Chouldechova, A., & Roth, A. (2018). The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 797-806).

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023.*

Cowgill, B. (2018). Bias and productivity in humans and algorithms: Theory and evidence from resume screening. Columbia Business School, Columbia University, 29.

Cowgill, B., Dell'Acqua, F., & Matz, S. (2020). The managerial effects of algorithmic fairness activism. In AEA Papers & Proceedings (Vol. 110).

Cowgill, B., & Tucker, C. (2017). Algorithmic bias: A counterfactual perspective. NSF Trustworthy Algorithms.

Crawford, K. (2013). The hidden biases in big data. *Harvard Business Review*, retrieved from https://hbr.org/2013/04/the-hidden-biases-in-big-data.

De-Arteaga, M., Romanov, A., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi. K., & Kalai, A. T. (2019). Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 120-128).

Dobbie, W., Liberman, A., Paravisini, D., & Pathania, V. (2018). Measuring bias in consumer lending (No. w24953). National Bureau of Economic Research.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (pp. 214-226).

Elzayn, H., Jabbari, S., Jung, C., Kearns, M., Neel, S., Roth, A., & Schutzman, Z. (2019). Fair algorithms for learning in allocation problems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 170-179).

Ezrachi, A., & Stucke, M. E. (2016). Virtual competition, *Journal of European Competition Law & Practice*, 7(9), 585–586

Executive Office of the President, Munoz, C., Director, D. P. C., Megan (US Chief Technology Officer Smith (Office of Science and Technology Policy)), & DJ (Deputy Chief Technology Officer for Data Policy and Chief Data Scientist Patil (Office of Science and Technology Policy)). (2016). Big data: A report on algorithmic systems, opportunity, and civil rights. Executive Office of the President.

Dastin, J. (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. Retreived from https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 259-268).

Fu, R., Aseri, M., Singh, P. V., & Srinivasan, K. (2019). 'Un'fair machine learning algorithms. *Available at SSRN 3408275*.

Fu, R., Huang, Y., & Singh, P. V. (2019). Crowds, lending, machines, and bias. *Available at SSRN 3206027*.

Fu, R., Huang, Y., Singh, P.V., & Srinivasan, K. (2020) When algorithms promote inequality: A structural analysis of the impact of Zillow's Zestimate on housing market. Working paper.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2018). Predictably unequal? The effects of machine learning on credit markets. *Available at SSRN 3072038*.

Ford, R. A., Price, W., & Nicholson, I. I. (2016). Privacy and accountability in black-box medicine. *Michigan Telecommunications and Technology Law Review*, 23(1).

Greiner, D. J., & Rubin, D. B. (2011). Causal effects of perceived immutable characteristics. *Review of Economics and Statistics*, *93*(3), 775-785.

Hardt, M. (2014). How big data is unfair: Understanding sources of unfairness in data driven decision making. *Unpublished paper.(medium. com/@mrtz/how-big-datais-unfair-9aa544d739de)*.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 3323-3331).

Hooker, J. N., & Williams, H. P. (2012). Combining equity and utilitarianism in a mathematical programming model. *Management Science*, *58*(9), 1682-1693.

Ilvento, C., Jagadeesan, M., & Chawla, S. (2020). Multi-category fairness in sponsored search auctions. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 348-358).

Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., & Roth, A. (2017). Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning* (Volume 70, pp. 1617-1626).

Jacobs, A. Z., Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). The meaning and measurement of bias: Lessons from natural language processing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 706-706).

Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines. arXiv preprint arXiv:1906.11668.

Johndrow, J. E., & Lum, K. (2017). An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*.

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 325-333).

Jung, J., Corbett-Davies, S., Shroff, R., & Goel, S. (2018). Omitted and included variable bias in tests for disparate impact. *arXiv preprint arXiv:1809.05651*.

Kamiran, F., Calders, T., & Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In *2010 IEEE International Conference on Data Mining* (pp. 869-874). IEEE.

Kilbertus, N., Carulla, M. R., Parascandolo, G., Hardt, M., Janzing, D., & Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 656-666).

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference, ITCS.*

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics*, *133*(1), 237-293.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic fairness. In *AEA Papers and Proceedings* (Vol. 108, pp. 22-27).

Kroll, J. A., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*, 633-705.

Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 4069-4079).*

Lakkaraju, H., Kleinberg, J., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). The selective labels problem: Evaluating algorithmic predictions in the presence of unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 275-284)*.

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science*, 65(7), 2966-2981.

Lange, J. E., Blackman, K. O., & Johnson, M. B. (2001). Speed violation survey of the New Jersey Turnpike. Public Services Research Institute.

Larson, J., Mattu, S., Kirchner, L., & Angwin. J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*.

Lee, N. T., Resnick, P., & Barton, G. (2019). Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms. Center for Technology Innovation. Brookings. Retrieved from https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/

Manyika, J., Silberg, J., & Presten, B. (2019) What do we do about the biases in AI? *Harvard Business Review*, retrieved from https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai.

McConnell, E. H., & Scheidegger, A. R. (2001). Race and speeding citations: Comparing speeding citations issued by air traffic officers with those issued by ground traffic officers. In *Annual Meeting of the Academy of Criminal Justice Sciences, Washington, DC*.

Miklós-Thal, J., & Tucker, C. (2019). Collusion by algorithm: Does better demand prediction facilitate coordination between sellers? *Management Science,* 65(4), 1552-1561.

Nabi, R., & Shpitser, I. (2018). Fair inference on outcomes. *In Thirty-Second AAAI Conference on Artificial Intelligence* (pp. 1931-1940).

Nkonde, N., 2019, Is AI bias a corporate social responsibility issue? *Harvard Business Review*, retrieved from https://hbr.org/2019/11/is-ai-bias-a-corporate-social-responsibility-issue.

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

Pedreshi, D., Ruggieri, S., & Turini, F. (2008). Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 560-568).

Simoiu, C., Corbett-Davies, S., & Goel, S. (2017). The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, *11*(3), 1193-1216.

Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)* (pp. 1-7).

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, *7*(2), 76-99.

Wang, Q., Huang, Y., Jasin, S., & Singh, P.V. (2020) Algorithmic transparency with strategic users. Working paper.

Yao, S., & Huang, B. (2017). Beyond parity: Fairness objectives for collaborative filtering. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 2925-2934).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning*, 28(3), 325-333.